به‌کارگیری داده‌کاوی برای پیشنهاد پرسش در نظام‌های بازیابی اطلاعات

مهدی زینالی تازه کندی*[1]، محسن نوکاریزی[2]

مطالعات دانش‌شناسی
سال هفتم، شماره ۲۳، تابستان ۹۹، ص ۱ تا ۱۸

تاریخ دریافت: ۹۸/۰۳/۱۵
تاریخ پذیرش: ۹۸/۱۲/۱۰

چکیده

داده‌کاوی به مفهوم آشکارسازی الگوهای موجود در حجم انبوه داده‌هاست که در بسیاری از رشته‌ها به کار گرفته شده است. در رشته علم اطلاعات و دانش‌شناسی به‌ویژه در بازیابی اطلاعات نیز می‌توان از آن بهره برد. در بازیابی اطلاعات ابتدا پارادایم نظام‌گرا و سپس پارادایم کاربرگرا مطرح شده است که در پارادایم دوم به نیاز اطلاعاتی توجه شده است. در پارادایم دوم، ورود پرسش‌های نامناسب از سوی کاربران، دلیل اصلی عدم بازیابی مدارک مرتبط تلقی می‌شود. ازاین‌رو، یکی از مباحث اصلی این پارادایم، پیشنهاد و بسط پرسش مناسب در نظام بازیابی اطلاعات است که می‌توان از روش‌های داده‌کاوی برای آن استفاده کرد. چهار روش مهم برای پیشنهاد پرسش جهت تقویت نظام توصیه‌گر وجود دارد. قاعده سری زمانی یکی ازاین‌روش‌هاست که به فراوانی پرسش در واحد زمانی خاص می‌پردازد. یکی دیگر از روش‌ها، قانون همایندی است که به وابستگی و تداعی پرسش‌ها توجه دارد. در روش قانون همایندی همراه با فاصله لون اشتاین، افزون بر توجه به وابستگی و تداعی پرسش‌ها به ترتیب واژه‌های پرسش نیز توجه می‌شود. به‌هرحال، در هر سه روش یادشده، از فایل ثبت رخداد استفاده می‌شود؛ درحالی‌که در نظریه احتمالاتی از واژه‌های مدارک جهت ترمیم شکاف واژگانی بین پرسش و مدارک استفاده می‌شود. درنهایت به نظر می‌رسد، به‌کارگیری روش‌های یادشده به‌ویژه روش احتمالی در پیشنهاد پرسش منجر به نتایج مناسب‌تری شود.

واژه‌های کلیدی: سامانه توصیه‌گر، فاصله لون اشتاین ، قاعده سری زمانی، قانون همایندی، نظریه احتمالی

۱.* دانشجوی دکتری علم اطلاعات و دانش شناسی، گروه علم اطلاعات و دانش شناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه فردوسی مشهد، مشهد، ایران. Ma.zeynali@mail.um.ac.ir

۲. دانشیار رشته علم اطلاعات و دانش شناسی، گروه علم اطلاعات و دانش شناسی، دانشکده علوم تربیتی و روان‌شناسی، دانشگاه فردوسی مشهد، مشهد، ایران. mnowkarizi@um.ac.ir

مقدمه

امروزه وب به مخزن عظیمی از اطلاعات تبدیل‌شده است که تقریباً همه موضوعات موردعلاقه کاربران انسانی را پوشش می‌دهد و این امر موجب محبوبیت وب برای کاربران شده است. با افزایش محبوبیت جهانی وب، مقدار حجیمی از داده‌ها در سرورهای وب در فایل‌های تراکنش ذخیره می‌شود. در این فایل‌ها، تمامی رخدادهای انجام‌شده توسط کاربران ثبت می‌گردد. ازاین‌رو، در سال‌های اخیر به دلیل دسترس‌پذیر بودن حجم انبوه داده‌ها توجه‌ها به داده‌کاوی معطوف شده است و آن به‌عنوان یکی از پیشرفت‌های اخیر محسوب می‌شود.

در سال‌های ۱۹۸۹ و ۱۹۹۱ کارگاه‌های اکتشاف دانش از پایگاه داده‌ها توسط پیاتتسکی[1] و همکارانش برگزار شد (حیاتی، صادقی مجرد و جعفری، ۱۳۸۹). در داده‌کاوی به بررسی و تجزیه‌وتحلیل مقادیر عظیمی از داده‌ها به‌منظور کشف الگوها و قوانین معنادار پرداخته می‌شود. داده‌کاوی را می‌توان استخراج اطلاعات و مفاهیم از پایگاه داده‌ها دانست که به‌صورت نیمه‌خودکار تغییرات، وابستگی‌ها و الگوها را کشف می‌کند و برای انجام این کار از علم آمار بهره می‌برد (رحمانی و زین‌العابدینی، ۱۳۹۴). می‌توان از داده‌کاوی در زمینه‌های مختلف استفاده کرد. در این رابطه پژوهشگران مختلف به کاربرد داده‌کاوی در حوزه‌های مختلف پرداخته‌اند که می‌توان به کاربرد داده‌کاوی در مطالعات اجتماعی (مرادی و قاسمی، ۱۳۹۱)، مدیریت (توکلی، مرتضوی، کاهانی و حسینی، ۱۳۸۹)، سلامت (قادر پور، ۱۳۹۶)، مهندسی (بدیعی و غضنفری، ۱۳۹۶) اشاره کرد. در رشته علم اطلاعات و دانش‌شناسی نیز داده‌کاوی کاربردهای فراوان دارد. در این راستا، پژوهشگران مختلف تلاش کرده‌اند تا مفهوم داده‌کاوی و کاربرد آن را تشریح کنند که می‌توان به چندین اثر نظیر صراف‌زاده و حاضری (۱۳۸۰)، جعفرپور (۱۳۹۰) و حیاتی و همکاران (۱۳۸۹) اشاره کرد.

بازیابی اطلاعات یک شاخه از علم اطلاعات و دانش‌شناسی است که به بازیابی اطلاعات مرتبط می‌پردازد. وقتی یک شخص به اطلاعاتی برای حل یک مشکل یا مسئله‌ای نیاز دارد؛

1. Piatetsky

از طریق ابزارهای مختلف به جستجو می‌پردازد تا با استفاده از اطلاعات مرتبط، نیاز خویش را مرتفع کند. در حالت کلی، بازیابی اطلاعات را می‌توان دارای دو بعد دانست که بعد اول مربوط به بازنمون مدارک و بعد دوم مربوط به نیاز اطلاعاتی است (اینگورسن[1]، ۱۳۸۹). در مورد عدم بازیابی اطلاعات مرتبط دو پارادایم وجود دارد (کنت و لانکو[2]، ۱۹۶۸). منظور از پارادایم، سرمشق و الگوی مسلط و چهارچوب فکری و فرهنگی است که مجموعه‌ای از الگوها و نظریه‌ها را برای یک گروه یا یک جامعه شکل می‌دهد. در پارادایم اول به نظام و بازنمون مدرک و در پارادایم دوم به کاربر و بازنمون نیاز اطلاعاتی توجه شده است. البته لازم به یادآوری است که توجه به هر دو پارادایم توسط پژوهشگران بازیابی اطلاعات نظیر کاپورا و یورلند[3] (۲۰۰۰)، باد[4] (۲۰۰۴)، ثورنلی و گیب[5] (۲۰۰۷)، ساراسویک[6] (۲۰۰۷، ۲۰۱۲) و یورلند (۲۰۱۰) ضروری شمرده شده است. اگر هرکدام از این دو پارادایم نادیده گرفته شود، رسیدن به اهداف دور از انتظار خواهد بود.

همان‌گونه که پیش‌تر ذکر شد، محبوبیت وب و نظام‌های بازیابی اطلاعات و درنتیجه ثبت داده‌های حجیم در فایل تراکنش‌ها موجب رشد و ترقی داده‌کاوی شد و اکنون خود به یکی از گرایش‌های دانشگاهی تبدیل شده است. اگرچه وب و نظام‌های بازیابی اطلاعات موجب پیدایش و پیشرفت داده‌کاوی شده است، اما می‌توان از داده‌کاوی به‌منظور تقویت نظام‌های بازیابی اطلاعات نیز استفاده کرد؛ اما جای خالی چنین آثاری که کاربرد داده‌کاوی را در این رابطه نشان دهد، احساس می‌شود. البته به دلیل جدید بودن موضوع داده‌کاوی این امر دور از ذهن به نظر نمی‌رسد، اما نیاز به تلاش‌های بیشتری به‌منظور بسط و کاربرد مفاهیم و روش‌های داده‌کاوی در حوزه ما احساس می‌شود. ازاین‌رو، در این مقاله تلاش شده است به گوشه‌ای از کاربرد روش‌های داده‌کاوی در رابطه با رشته علم اطلاعات و دانش‌شناسی به‌ویژه بازیابی اطلاعات پرداخته شود. برای رسیدن به این امر، ابتدا به‌صورت خلاصه به

1. Ingwersen
2. Kent and Lancou
3. Capurro and Hjørland
4. Budd
5. Thornley and Gibb
6. Saracevic

مباحث پیرامون بازیابی اطلاعات پرداخته شده است تا جایگاه روش‌های داده‌کاوی در بازیابی اطلاعات مشخص شود. سپس چهار روش داده‌کاوی که از آن می‌توان جهت تقویت نظام توصیه‌گر و پیشنهاد و بسط پرسش‌های کاربران استفاده کرد، تشریح شده است.

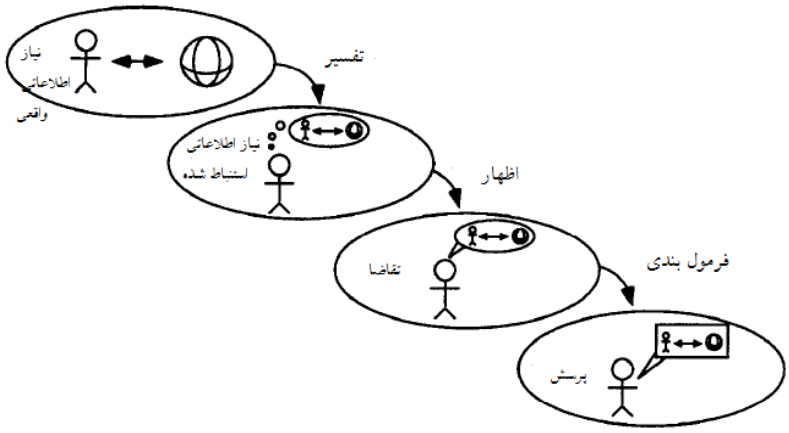
بازیابی اطلاعات. به‌محض این‌که اولین نظام رایانه‌ای بازیابی اطلاعات در سال ۱۹۵۰ ایجاد شد، مشکلات آن‌ها نیز با استفاده کاربران آشکار گردید. لذا این سؤال مطرح شد که چه عاملی موجب بازیابی منابع نامرتبط می‌شود (کنت و لانکو، ۱۹۶۸). در رابطه با به این سؤال دو پارادایم شکل گرفت. ازآنجایی‌که کتابداران از ابتدای شکل‌گیری این رشته در دانشگاه‌ها با بازنمون مدارک سروکار داشتند و عمده فعالیت آن‌ها را فهرست‌نویسی و سازمان‌دهی منابع اطلاعاتی شکل می‌داد، به نظر می‌رسد این امر موجب شد تا نخستین اندیشه‌ای که در ذهن آن‌ها در مورد بازیابی منابع نامرتبط در نظام‌های رایانه‌ای نیز نقش ببندد، عدم بازنمون صحیح منابع اطلاعاتی باشد. به‌این‌ترتیب، نخستین پارادایم بدین گونه شکل گرفت که دست نیافتن به اطلاعات مرتبط ناشی از عدم بازنمون صحیح مدارک است؛ البته فیدل[1] (۱۹۹۸) خاطرنشان می‌کند که تسلط رویکرد کمی بر پژوهش نیز بر این امر تأثیرگذار بوده است. درنتیجه، عمده پژوهش‌های اولیه در این رابطه مبتنی بر این پارادایم بود که می‌توان به اجلاس ترک[2] و آزمون کران فیلد[3] اشاره کرد. در برابر پارادایم نخست که در آن بازنمون مدارک موردتوجه قرار می‌گرفت، دروین و نیلان[4] (۱۹۶۸) تأکید کردند که باید به کاربر نیز توجه شود. در این رابطه ویلسون[5] (۲۰۰۰) اشاره می‌کند که انتقاد طرفداران رویکرد کیفی نسبت به رویکرد کمی در حوزه روش پژوهش بر پژوهشگران رشته علم اطلاعات و دانش‌شناسی نیز تأثیرگذار بوده است. در همین راستا فیدل (۲۰۰۸) نیز اعتقاد دارد که ظهور رویکرد کیفی در حوزه روش پژوهش موجب شکل‌گیری پارادایم کیفی در بازیابی اطلاعات شد. طرفداران این پارادایم، علت عدم دستیابی به اطلاعات مرتبط را ناشی

1. Fidel
2. Text Rerneval Conference
3. and Cranfield tests
4. Dervin and Nilan
5. Wilson

از عدم درک صحیح نیاز اطلاعاتی و تدوین و ارائه آن به ابزارهای جستجو می‌دانند (کنت و لانکو، ۱۹۶۸).

نیاز اطلاعاتی. نیاز اطلاعاتی یکی از مفاهیم اساسی در علم اطلاعات و دانش‌شناسی است که پژوهشگران مختلف نظیر درر[1] (۱۹۸۳)، تیمینز[2] (۲۰۰۶)، کول[3] (۲۰۱۱) به آن پرداخته‌اند. در واژه‌نامه برخط اودلیس[4] نوشته ریتز[5] (۲۰۱۹) نیاز اطلاعاتی، شکاف در دانش شخصی تعریف شده است وقتی که در سطح آگاهانه به‌عنوان یک درخواست یا تقاضا تجربه می‌شود و درنتیجه منجر به انجام جستجو برای یافتن پاسخ می‌شود. در همین راستا، بر اساس نظر میراندا و تاراپانوف[6] (۲۰۰۸) نیاز اطلاعاتی به حالت یا فرایندی اطلاق می‌شود که وقتی در فرد آغاز می‌شود که وی دریابد بین اطلاعات و دانش موجود وی و اطلاعات لازم برای حل یک مسئله یا مشکل شکاف وجود دارد. پس با توجه به تعاریف یادشده مشخص است که نیاز اطلاعاتی زمانی در فرد ایجاد می‌شود که فرد بداند که نمی‌داند. بعدازاین که فرد به این مرحله رسید، وی باید با استفاده از دانش خویش، ناآگاهی خود را در قالب کلمات برای انجام جستجو اظهار کند و این خود مسئلهٔ پیچیده و دشواری است. در همین رابطه میزارو[7] (۱۹۹۸) به مراحل تبدیل نیاز اطلاعاتی به پرسش پرداخته است که در ادامه (شکل ۱) به آن اشاره شده است.

1. Derr
2. Timmins
3. Cole
4. https://www.abc-clio.com/ODLIS/odlis_i.aspx
5. Reitz
6. Miranda and Tarapanoff
7. Mizzaro



شکل ۱. نیاز اطلاعاتی واقعی، نیاز اطلاعاتی درک شده، تقاضا و پرسش (اقتباس از: میزارو، ۱۹۹۸)

همان‌گونه که در شکل ۱ مشاهده می‌شود، چه‌بسا کاربر نیاز اطلاعاتی خود را به‌درستی استنباط نکند. افزون بر این، بیان نیاز درک شده و تبدیل آن به پرسش نیز خود مقوله دیگری است که باید در پارادایم کاربر به آن توجه شود. برای مشخص شدن این موضوع، کاربری را فرض کنید که نیاز اطلاعاتی الف را دارد. او نیاز اطلاعاتی را تبدیل به پرسش کرده و پرسش را در جعبه جستجوی موتور جستجو تایپ می‌کند. او نیاز اطلاعاتی دارد اما نمی‌داند که با چه کلماتی، پرسش خود را تدوین کند. ازآنجایی‌که مدارک به‌وسیله موتورهای جستجو نمایه‌سازی می‌شود و برای کاربر مشاهده‌پذیر نیست، اغلب واژگانی توسط کاربران برای انجام جستجو انتخاب می‌شود که به بازیابی مدارک موردنظر ختم نمی‌شود و این امر ناشی از شکاف بین فضای واژگان پرسش کاربران و فضای واژگان مدارک است (بهاشیا، معجوم دار، میترا[1]، ۲۰۱۱). برای توجه به پارادایم کاربری و نیز ازآنجایی‌که بیشتر کاربران موتورهای جستجو آشنایی کافی با موتورهای جستجو نداشته و برای تدوین پرسش‌ها آموزش لازم را ندیده‌اند (شی و یانگ[2]، ۲۰۰۷)، ضرورت آشنایی با فنون جستجویی وبی برای حل این مسئله احساس می‌شود. نظام‌های بازیابی اطلاعات برای کمک به کاربران در

1. Bhatia, Majumdar & Mitra
2. Shi andYang

این راستا می‌توانند پرسش‌های مشابه به پرسش‌های کاربران را پیشنهاد کنند (بهاشیا و همکاران، ۲۰۱۱). پیشنهاد پرسش عبارت است از یافتن پرسش‌های مرتبط به پرسش اصلی کاربر که توسط وی وارد شده است. برای نمونه، وقتی کاربر پرسش «هواپیمایی ایران» را وارد موتور جستجو می‌کند، موتور جستجو پرسش‌هایی نظیر «بلیط هواپیما»، «بلیط برخط هواپیما»، «رزرو هواپیمایی ایران» را به او پیشنهاد می‌کند (ویندلی و اوزاکان[1]، ۲۰۱۶). افزون بر این، پیشنهاد پرسش شامل پیشنهاد پرسش یا تایپ بخشی از پرسش توسط کاربر، اصلاح خطای املائی و نگارشی پرسش کاربران می‌شود.

متون مختلفی در مورد پیشنهاد پرسش به کاربران وجود دارد که در دو گروه قابل‌طبقه‌بندی است: گروه اول آثاری است که با استفاده از تحلیل فایل‌های ثبت رخداد کاربران پرسش‌هایی را پیشنهاد می‌دهند؛ گروه دوم آثاری است که نیاز به تحلیل فایل‌های ثبت رخداد ندارد، بلکه از تحلیل محتوای وب‌سایت‌ها و مدارک بازیابی شده استفاده می‌کند. افزون بر این، در برخی آثار نیز تلاش می‌شود با درک بافت کاربر، پرسش‌هایی را پیشنهاد دهند. در این رابطه گفتنی است که اطلاعات کاربران، نظیر جنسیت، سن، نام کاربری، نشانی IP و پرسش‌های قبلی او، به منزلهٔ بافت تلقی می‌شود (ویندلی و اوزاکان، ۲۰۱۶). ازآنجایی‌که توجه به پارادایم کاربری در رشته علم اطلاعات و دانش‌شناسی از اهمیت ویژه‌ای برخوردار است و درنهایت موجب تقویت نظام‌های بازیابی اطلاعات برای بازیابی منابع مرتبط‌تر می‌شود، در ادامه به روش‌های مختلف داده‌کاوی که موجب کمک به درک صحیح نیاز اطلاعاتی و ورود پرسش مناسب به نظام‌های بازیابی اطلاعات می‌شود، اشاره شده است.

روش‌های داده‌کاوی. امروزه داده‌های زیادی در موتورهای کاوش ثبت و ذخیره می‌شود که شناخت الگوهای بین آن‌ها موجب رسیدن به دانش در آن حوزه می‌شود. داده‌کاوی را می‌توان کشف الگوهای موجود در دادهای انبوه دانست که در آن از فنون آماری، ریاضی، هوش مصنوعی، یادگیری ماشین و نظیر آن استفاده می‌شود (مرادی و قاسمی، ۱۳۹۱). در

1. Vidinli & Ozcan

داده‌کاوی از روش‌های مختلفی نظیر درختواره تصمیم[1]، قواعد همایندی[2]، نظریه احتمالات[3] و نظیر آن استفاده می‌شود که در این مقاله، چهار نوع روش مهم برای پیشنهاد پرسش در موتورهای کاوش شناسایی شد و در ادامه به هر یک از آن‌ها پرداخته شده است.

قاعده سری زمانی[4] که همبستگی زمانی[5] نیز نامیده می‌شود، بر اساس نظر بروکول و دیویس[6] (۲۰۱۶) به مجموعه مشاهدات ثبت‌شده‌ای اشاره دارد که هر مشاهده در زمان خاصی ثبت شده است و منظور از سری زمانی مجموعه‌ای از داده‌های آماری است که در فواصل زمانی مشخص و منظمی گردآوری شده باشند. عمادالله[7] (۲۰۱۳) توالی یا یک رشته از متغیرهای تصادفی نمایه شده بر اساس زمان را یک فرآیند دارای تغییر در زمان‌های مختلف[8] یا سری زمانی نامیده است. از این قاعده می‌توان برای تقویت نظام توصیه‌گر استفاده کرد. به‌منظور به‌کارگیری این روش باید از همبستگی زمانی برای پیشنهاد پرسش‌های مرتبط بهره برد. به بیانی دیگر، اگر محبوبیت دو پرسش در طول زمان مشابه باشد؛ آنگاه آن دو پرسش مرتبط تلقی می‌شود. با استفاده ازاین‌روش می‌توان فهمید که چرا یک پرسش در زمانی خاص موردتوجه قرار گرفته است؛ برای مثال «شکلات» پرسشی است که در ماه فوریه بیشتر موردتوجه کاربران قرار می‌گیرد؛ همچنین می‌توان پیش‌بینی کرد که پرسش‌های مرتبط با جشن عید نوروز در زمان جشن عید نوروز موردتوجه کاربران خواهد بود. در این راستا، چن و ایمورولیکا[9] (۲۰۰۵) با استفاده از فایل تراکنش موتور جستجوی ام‌اس‌ان پرسش‌های مرتبط به هم را بر اساس قاعده یاد شده به دست آوردند و در انتها فرمول زیر را برای استخراج پرسش‌های مرتبط ارائه کردند. طبق نظر آن‌ها، برای محاسبه همبستگی دو پرسش p و q می‌توان از ضریب همبستگی تابع فراوانی[10] آن‌ها به شرح ذیل استفاده کرد.

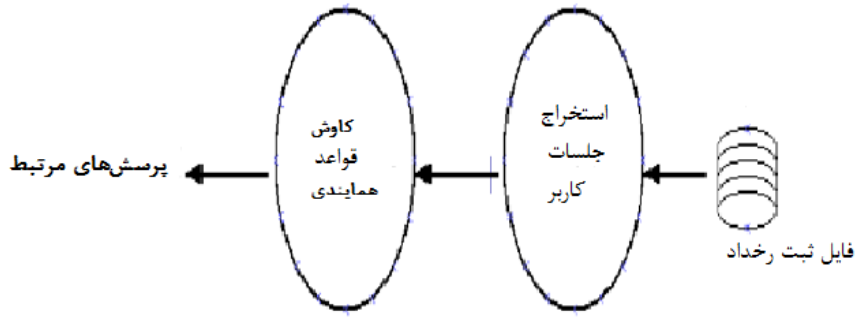
1. decision tree
2. association rule
3. probability theory
4. time series
5. temporal correlation
6. Brockwell and Davis
7. Imdad Ullah
8. stochastic
9. Chien & Immorlica
10. correlation coefficient of their frequency functions

$$\frac{1}{d} \sum \left(\frac{x_{p_i} - \mu(x_p)}{\sigma(x_p)}\right) \left(\frac{x_{q_i} - \mu(x_q)}{\sigma(x_q}\right)$$

در این فرمول منظور از x_{q_i} فراوانی پرسش q در زمان i یا i امین واحد زمانی است. همچنین می‌توان آن را $\frac{n_{q_i}}{N_i}$ در نظر گرفت که n_{q_i} تعداد رویدادهای پرسش q در زمان i یا i امین واحد زمانی و N_i تعداد کل پرسش‌ها در زمان i یا i امین واحد زمانی است.پس x_{q_i}، تابع فراوانی پرسش q، $M(x_q)$، میانگین فراوانی پرسش q و $\sigma(x_q)$، انحراف معیار فراوانی پرسش q را نشان می‌دهد. نتیجه حاصل از فرمول یاد شده، ضریب همبستگی دو پرسش p و q را نشان می‌دهد که عددی بین ۱- و ۱+ خواهد بود. عدد منفی یک، نشان‌دهندهٔ رابطه کامل غیرمستقیم و عدد مثبت یک، نشان‌دهندهٔ رابطه مثبت کامل است (چن و ایمورولیکا، ۲۰۰۵).

قاعده همایندی. در میان فنون داده‌کاوی، پژوهشگران به کشف الگوهای مکرر توجه خاصی نشان داده‌اند. همان‌گونه که از نام این قاعده پیداست؛ در آن به دنبال الگوهایی هستیم که به‌مراتب تکرار می‌شوند. در قوانین همایندی وابستگی عمده میان اقلام موجود در پایگاه داده یا فایل تراکنش مشخص می‌شوند، به‌نحوی‌که حضور برخی از اقلام بر حضور برخی از اقلام دیگر در فایل تراکنش دلالت دارد (نورانی، ستاری و مولاجو، ۱۳۹۵). به‌بیان‌دیگر، قاعده همایندی روابط و وابستگی‌های متقابل بین مجموعه عظیمی از اقلام داده‌ای را نشان می‌دهد (مرادی، منعم و مرادی، ۱۳۹۴). برای نمونه، سیاه‌وسفید؛ شب و روز؛ و بسته و باز به‌طورمعمول در کنار هم می‌آیند؛ یا کاربرد یکی دیگری را تداعی می‌کند. پس به نظر می‌رسد، عناوین «قاعده یا قانون همایندی» یا «قاعده یا قانون تداعی» مناسب‌تر از عنوان «قوانین انجمنی» باشد. ازاین‌رو، اگرچه در اغلب آثار نظیر شاهین و رضازاده (۱۳۹۰)، مرادی و همکاران (۱۳۹۴) و نورانی و همکاران (۱۳۹۵) از عنوان «قوانین انجمنی» استفاده شده است، نویسندگان مقاله حاضر از عنوان «قاعده همایندی» استفاده کرده‌اند. به‌هرحال، پیدا کردن چنین قواعدی می‌تواند در علم اطلاعات و دانش‌شناسی موردتوجه بوده و کاربردهای متفاوتی داشته باشد؛ مثلاً کشف روابط همایندی بین حجم عظیم تراکنش‌های رفتار کاربران

برای شخصی‌سازی مورداستفاده قرار گیرد. افزون بر این، قاعده همایندی در تقویت سامانه توصیه‌گر نظام‌های بازیابی اطلاعات نیز کاربرد دارد. در همین راستا، فونسکا، مورا، گلگهر و زیوانی[1] (۲۰۰۳) از پرسش‌های وارد شده به موتور جستجوی برزیلی فرجودور[2] استفاده و روش خودکاری را برای پیشنهاد پرسش به‌منظور کمک به کاربران ارائه کردند. آن‌ها اطلاعات موردنیاز خود را از فایل ثبت رخداد— پرسش‌های قبلاً وارد شده— استخراج کرده و قوانین همایندی را مشخص کردند. این پژوهشگران در پژوهش خود از بیش از ۲.۳ میلیون پرسش وارد شده به موتور جستجوی فرجودور استفاده و روشی را ارائه کردند که دقت این روش در پنج پرسش پیشنهادی ۹۰.۵ درصد گزارش شد. روش این پژوهشگران برای شناسایی پرسش‌های مرتبط دومرحله‌ای بود که در شکل ۲ نشان داده شده است.



شکل ۲. تشخیص پرسش‌های مرتبط (اقتباس از: فونسکا و همکاران، ۲۰۰۳)

همان‌گونه که در شکل ۲ مشاهده می‌شود، در مرحله اول فایل ثبت رخداد موتورهای جستجو تحلیل‌شده و جلسات کاربر استخراج می‌شد. در مرحله دوم قوانین همایندی از مجموعه جلسات کاربر واکاوی شده و پرسش‌های مرتبط استخراج می‌شد.

قاعده همایندی همراه بافاصله لون اشتاین. در پیشنهاد پرسش با استفاده از قاعده همایندی، از فایل ثبت پرسش‌های کاربران در بازه زمانی ده‌دقیقه‌ای استفاده می‌شود و ممکن است این خطا رخ دهد که موضوع پرسش در این بازه زمانی تغییر کند؛ شی و یانگ (۲۰۰۷)

1. Fonseca, Golgher, Moura, & Ziviani
2. Farejador

برای غلبه بر این مشکل، روش دیگری متشکل از سه مرحله برای استخراج پرسش‌های مرتبط پیشنهاد دادند که در ادامه به آن اشاره شده است.

مرحله نخست: استخراج داده‌های جلسات کاربر از فایل ثبت رخداد مربوط به پرسش

در این مرحله به داده‌های اصلی سوابق پرسش موتورهای جستجو پرداخته می‌شود. جلسات کاربر با استفاده از گزارش‌های مربوط به پرسش استخراج می‌شود. سوابق مربوط به پرسش متعلق به همان کاربری است که با پروتکل منحصربه‌فرد اینترنت[1] خود مشخص شده است. درصورتی‌که از رایانه‌های مشترک عمومی استفاده شود، فرض ما تضعیف می‌شود. باوجوداین، روش ما بازهم می‌تواند اثرات منفی احتمالی ناشی از سردرگمی جلسات کاربران را خنثی کند؛ زیرا شناسایی پرسش‌های مرتبط به‌صورت آماری و مبتنی بر تراکنش[2] پرسش‌های استخراجی است نه جلسات کاربران.

مرحله دوم: تقسیم جلسات کاربر به تراکنش‌های پرسش

در این مرحله جلسات استخراج‌شده کاربر به‌عنوان ورودی پذیرفته و سپس هر یک از آن‌ها به‌صورت مناسب به تراکنش‌های پرسش تقسیم می‌شود. در ادامه تراکنش‌های پرسش تقسیم‌شده با هم جمع می‌شوند. تا پایان این مرحله، از هویت کاربر و نشان‌های زمانی تراکنش پرسش استفاده نمی‌شود.

مرحله سوم: کشف پرسش‌های مرتبط از تراکنش‌های پرسش

در این مرحله، اگر برخی از شرایط مهیا باشد، پرسش اولیهٔ وارد شده توسط کاربر به‌صورت مستقیم به‌عنوان درونداد این روش در نظر گرفته می‌شود. تراکنش‌های پرسش از گنجینه تراکنش‌ها جمع می‌شود و سپس ارتباط بین پرسش ورودی و هر پرسش دیگر دارای محدودیت‌های از پیش تعریف شده محاسبه می‌شود. معیارها و محدودیت‌های از پیش تعریف شده ممکن است مبتنی بر فراوانی خام پرسش‌ها باشد اما فقط محدود به آن نیست.

1. internet protocol
2. transaction

شی و یانگ (۲۰۰۷) در پژوهش خود افزون بر استفاده از قاعده همایندی، از فاصله لون اشتاین[1] استفاده کردند. ازآنجایی‌که برای استخراج پرسش‌های مرتبط، نیاز به سنجش ارتباط (شباهت و نزدیکی) دو پرسش است؛ پس اگر از عدد یک، فاصله به‌دست‌آمده کم شود، نتیجه حاصل شباهت و نزدیکی دو پرسش را نشان خواهد داد. بدین ترتیب، شباهت و ارتباط بین دو پرسش بر اساس فاصله لون اشتاین از فرمول زیر قابل‌محاسبه است (شی و یانگ،۲۰۰۷):

$$similarity_{Levenshtein} = 1 - \frac{Levenshtein\ distance(q_1,\ q_2)}{\max\ (wn(q_1),wn(q_2))}$$

در این فرمول، منظور از $similarity_{Levenshtein}$، ارتباط دو پرسش؛ $Levenshtein\ distance\ (q_1,\ q_2)$، فاصله لون اشتاین پرسش۱ و ۲؛ و $wn(q_1)$، تعداد کلمات پرسش ۱ است. در این فرمول عدد ۱ بیشترین ارتباط و صفر کم‌ترین ارتباط بین دو پرسش را نشان می‌دهد. پژوهشگران یاد شده به‌منظور تعیین دقت روش خویش، آن را با دو روش قبلی مقایسه کردند. آن‌ها اظهار داشتند که دقت قاعده همایندی بیشتر از قاعده سری زمانی است. درنتیجه روش پیشنهادی آن‌ها که استفاده از قاعده همایندی همراه بافاصله لون اشتاین بود، دقت بالایی در مقایسه با دو روش قبلی داشت.

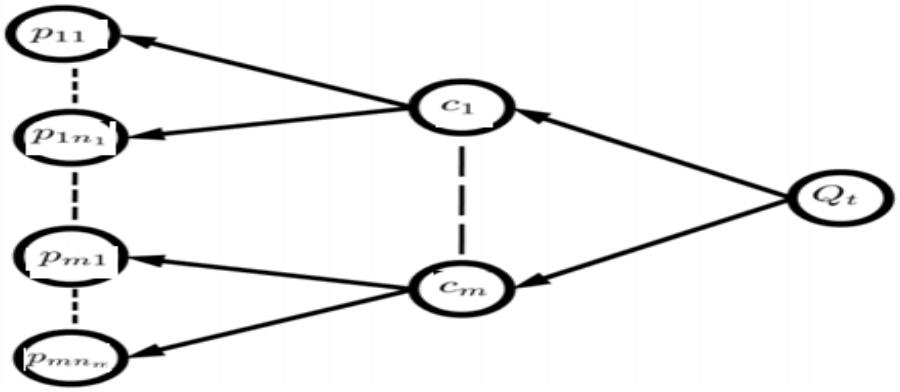
نظریه احتمالات: استفاده از مدارک

در سه روش قبلی از فایل ثبت رخداد موتورهای جستجو استفاده می‌شد؛ لذا در این روش‌ها، پرسش‌های پیشنهادی برخاسته از پرسش‌های وارد شده بود. ازاین‌رو، امکان عدم‌جامعیت پرسش‌های پیشنهادی وجود داشت. بر همین اساس چو، ون، نی و ما[2] (۲۰۰۲) و بهاشیا و همکاران (۲۰۱۱) برای حل این مسئله، استفاده از واژه‌های مدارک؛ و برای رتبه‌بندی پرسش‌های پیشنهادی، استفاده از نظریه احتمالات را پیشنهاد دادند.

مفهوم احتمال در مورد ارتباط یا پیوند دو چیز به کار می‌رود، به این معنی که ارتباط یا پیوند آن‌ها به صورتی است که حضور، شکل، وسعت و اهمیت هر یک وابسته به حضور،

1. Levenshtein distances
2. Cui, Wen, Nie & Ma

شکل و اهمیت دیگری است. نظریه احتمال علاوه بر توضیح پدیده‌های تصادفی به بررسی پدیده‌هایی می‌پردازد که لزوماً تصادفی نیستند؛ ولی با تکرار زیاد دفعات آزمایش نتایج از الگویی مشخص پیروی می‌کند، مثلاً در آزمایش پرتاب سکه با تکرار آزمایش می‌توان احتمال وقوع پدیده‌های مختلف را حدس زد (سیگموند[1]، ۲۰۱۹). در همین راستا، بهاشیا و همکاران (۲۰۱۱) برای کشف پرسش‌های مرتبط، به‌کارگیری نظریه احتمالات را پیشنهاد دادند که در شکل ۳ به آن اشاره شده است.



شکل ۳. استفاده از نظریه احتمالات برای پیشنهاد پرسش (اقتباس از بهاشیا و همکاران،۲۰۱۱)

در شکل ۳، q_t یک پرسش را نشان می‌دهد؛ امکان دارد این پرسش از چند واژه c_1 تا c_m تشکیل شده باشد. افزون بر این ممکن است که یک واژه، چندین واژه مترادف و هم‌معنا از p_{11} تا p_{n1} داشته باشد. این واژه‌های تشکیل‌دهندهٔ پرسش و واژه‌های هم‌معنا از مدارک استخراج و رتبه‌بندی پرسش‌های پیشنهادی با استفاده از نظریه احتمالات انجام می‌پذیرد.

بحث و نتیجه‌گیری

بازیابی اطلاعات را می‌توان شامل دو بعد مهم بازنمون مدرک و بازنمون نیاز اطلاعاتی دانست. به‌بیان‌دیگر، در هر نظام ذخیره و بازیابی، در یک‌سو با درک آغازین یا تجلیاتی ذهنی که در قالب متون مضبوط عینیت یافته‌اند و در سوی دیگر با پرسشگر یا کاربر روبرو

1. Siegmund

هستیم (ساراسویک، ۲۰۰۷). ازاین‌رو، بازیابی مدارک مرتبط به هر دو بعد یاد شده بستگی دارد.

مطالعات نشان داده است که توانایی کاربران در رسیدن به نتایج مطلوب ضعیف است که مهم‌ترین علت آن، ناتوانی کاربران در انتخاب واژه صحیح و عملگر مناسب در جستجوهاست (اسپینک، جانسون و اوزمولتو[1]، ۲۰۰۰؛ لوکاس و توپی[2]، ۲۰۰۵). لذا به نظر می‌رسد که توجه به بعد دوم فرایند بازیابی اطلاعات–بازنمون نیاز اطلاعاتی- نیاز به توجه بیشتری دارد. پیشنهاد کلیدواژه‌های مناسب و بسط جستجو یکی از ویژگی‌های مهم موتورهای کاوش است که در شکل دادن به رفتار اطلاع‌یابی کاربر مؤثر است و به او کمک می‌کند تا به نتایج بهتری دست یابد.

داده‌کاوی یکی از موضوعات جدیدی است که در سال‌های اخیر رونق یافته است و به کشف دانش از داده‌های عظیم می‌پردازد. به‌بیان‌دیگر الگوهای موجود و همیشگی و روابط داده‌ها را تجزیه‌وتحلیل می‌کند (رحمانی، زین‌العابدینی، ۱۳۹۴). ازآنجایی‌که پرسش‌های وارد شده به جعبه جستجوی موتورهای کاوش و سایر نظام‌های بازیابی اطلاعات و سایر فعالیت‌های کاربران در فایل تراکنش نظام ثبت و ذخیره می‌شود، لذا می‌توان در این حجم انبوه داده‌ها، با استفاده از داده‌کاوی به کشف و شناسایی الگوهای موجود بین داده‌ها پرداخت (حیاتی و همکاران، ۱۳۸۹). درنتیجه با به‌کارگیری الگوی موجود در فایل تراکنش به درک صحیح نیاز اطلاعاتی و تدوین مناسب پرسش توسط کاربران یاری کرد. در این راستا، روش‌های مختلفی در داده‌کاوی وجود دارد که می‌توان با استفاده از آن‌ها برای پیشنهاد پرسش به کاربران استفاده کرد که قاعده سری زمانی (چن و ایمورولیا، ۲۰۰۵)، قاعده همایندی (فونسکا و همکاران، ۲۰۰۳)، قاعده همایندی همراه بافاصله لون اشتاین (شی و یانگ، ۲۰۰۷) و نظریه احتمالات (چو و همکاران، ۲۰۰۲؛ بهاشیا و همکاران، ۲۰۱۱) ازجمله مهم‌ترین این روش‌هاست. در قاعده سری زمانی، پرسش‌های ثبت شده در فایل ثبت رخداد نظام‌های بازیابی اطلاعات در بازه‌های زمانی مساوی و منظم گردآوری و تحلیل

1. Spink, Jansen & Ozmultu
2. Lucas & Topi

می‌شود. در این روش می‌توان گفت که فقط به زمان توجه می‌شود. در قاعده همایندی افزون بر توجه به زمان، به وابستگی و تداعی پرسش‌های مختلف نیز توجه می‌شود؛ اما بازهم ممکن است که موضوع پرسش وارده توسط کاربران در بازه زمانی مشخص شده تغییر یابد. ازاین‌رو، شی و یانگ (۲۰۰۷) استفاده از فاصله لون اشتاین را همراه با قاعده همایندی پیشنهاد دادند. همین امر سبب می‌شود که به ترتیب واژه‌ها نیز توجه شود. اگرچه روش اخیر از دو روش قبلی تکامل‌یافته‌تر است، اما مسئله دیگری نیز در این رابطه وجود دارد و آن‌هم این است که مبنای هر سه روش یاد شده، فایل ثبت رخداد نظام بازیابی اطلاعات است. به‌بیان‌دیگر، پرسش‌هایی که پیش‌تر توسط کاربران به نظام‌های بازیابی اطلاعات وارد شده است مبنایی برای رسیدن به کل پرسش‌هاست. ازاین‌رو ممکن است پرسش‌های موجود در فایل ثبت رخداد جامع نباشند. همین عامل موجب پیشنهاد استفاده از نظریه احتمالات و استفاده از واژه‌های مدارک توسط بهاشیا و همکاران (۲۰۱۱) شد. درنهایت به نظر می‌رسد، برای پیشنهاد پرسش استفاده از نظریه احتمالات در مقایسه با قاعده سری زمانی، قاعده همایندی و قاعده همایندی به همراه فاصله لون اشتاین نتایج بهتری را به همراه داشته باشد.

منابع

اینگورسن، پیتر. (۱۳۸۹). *تعامل بازیابی اطلاعات*. ترجمه هاجر ستوده. تهران: کتابدار

باج پای، آر پی؛ دویودی، روپش کی. (۱۳۹۰). مروری بر کاربردهای داده‌کاوی در کتابداری و اطلاع‌رسانی. ترجمه اسماعیل جعفر پور. *کتاب ماه کلیات*، ۱۴(۹)، ۸۰-۸۵.

بدیعی، اقدس و غضنفری، مهدی. (۱۳۹۶). کاربرد داده‌کاوی در مهندسی تولید محصول از طراحی مفهومی تا تولید نهایی. *فصلنامه مدیریت زنجیره تأمین*،۵۷، ۴۵-۶۱.

پاتکار، ویک آن. (۱۳۸۰). *کاربردهای داده‌کاوی در کتابخانه‌ها و مؤسسات دانشگاهی*. ترجمه مریم صراف زاده و افسانه حاضری. ارتباط علمی، ۳(۵). بازیابی شده در https://www.unp.ir/article/university/sources-scientific/1582

توکلی، احمد؛ مرتضوی، سعید؛ کاهانی، محسن؛ و حسینی، زهرا. (۱۳۸۹). به‌کارگیری فرایند داده‌کاوی برای پیش‌بینی الگوهای رویگردانی مشتری در بیمه. *فصلنامه چشم‌انداز مدیریت بازرگانی*، ۹(۴)، ۴۱-۵۵.

حیاتی، زهیر؛ صادقی مجرد، مرجان؛ و جعفری، نیما. (۱۳۸۹). کشف مسیر حرکت کاربران اطلاعات الکترونیکی با استفاده از الگوریتم قوانین وابستگی در داده‌کاوی: مطالعه موردی وب‌سایت کتابخانه دانشگاه یو تی اس استرالیا. *فصلنامه کتابداری و اطلاع‌رسانی*، ۱۳(۱)، ۲۵۱-۲۸۳.

رحمانی، مهدی؛ و زین‌العابدینی، محسن. (۱۳۹۴). کاربردهای داده‌کاوی در علم اطلاعات و دانش شناسی. *فصلنامه مدیریت اطلاعات و دانش شناسی*، ۲(۳)، ۲۳-۳۲.

شاهین، آرش؛ و صالح زاده، رضا. (۱۳۹۰). طبقه‌بندی نیازهای مشتریان و تجزیه‌وتحلیل رفتار آن‌ها با استفاده از الگوی تلفیقی کانو و قوانین انجمنی. *فصلنامه تحقیقات بازاریابی نوین*، ۱ (۲)، ۱-۱۶.

قادر پور، نیلوفر. (۱۳۹۶). داده‌کاوی با الگوریتم داده در روند سلامت. *نشریه نخبگان علوم و مهندسی*، ۲(۱)، ۱۰۳-۱۰۹

مرادی، گلمراد؛ و قاسمی، وحید. (۱۳۹۱). تکنیک داده‌کاوی و کاربرد آن در مطالعات اجتماعی. *نشریه علوم اجتماعی*، ۹(۱)، ۱۵۷-۱۷۸.

نورانی، وحید؛ ستاری، محمدتقی؛ و مولاجو، امیر. (۱۳۹۵). روش ترکیبی درخت تصمیم و قوانین انجمنی در پیش‌بینی بلندمدت بارش. *مجله مدیریت آب و آبیاری*، ۶(۲)، ۳۳۰-۳۴۶.

References

Bhatia, S., Majumdar, D., & Mitra, P. (2011). Query suggestions in the absence of query logs. In Proceedings of the *34th international ACM SIGIR conference on Research and development in Information Retrieval* (pp. 795-804). ACM.

Borlund, P. (2003). The concept of relevance in IR. J*ournal of the American Society for information Science and Technology, 54* (10), 913-925.

Brockwell, P. J., & Davis, R. A. (2016). I*ntroduction to time series and forecasting*. springer.

Budd, J. M (2004). Relevance: Language, semantics, philosophy. *Library Trend, 52* (3).

Capurro, R., & Hjørland, B. (2003). The concept of information. *Annual Review of Information Science and Technology, 37*(1), 343-411.

Chien, S., & Immorlica, N. (2005). Semantic similarity between search engine queries using temporal correlation. In Proceedings of the 14*th international conference on World Wide Web* (pp. 2-11). ACM

Croft, W. B., Metzler, D., & Strohman, T. (2010). *Search engines: Information retrieval in practice* (Vol. 520): Addison-Wesley Reading

Cui, H., Wen, J. R., Nie, J. Y., & Ma, W. Y. (2002, May). Probabilistic query expansion using query logs. In *Proceedings of the 11th international conference on World Wide Web* (pp. 325-332). ACM.

Derr, R. L. (1983). A conceptual analysis of information need. *Information Processing & Management, 19*(5), 273-278.

Dervin, B. & Nilan, M. S. (1986). Information needs and use. *Annual Review of Information Sczence and Technology*, 21, 3-33

Fidel, R. (2008) Are we there yet?: Mixed methods research in library and information science. *Library & Information Science Research, 30*, 265-272.

Fidel, R (1993). Qualitative methods in information retrieval research. *Library and Information Science Research, 15*, 219-219.

Fonseca, B. M., Golgher, P. B., de Moura, E. S., & Ziviani, N. (2003, November). Using association rules to discover search engines related queries. In Web Congress, 2003. Proceedings. *First Latin American* (pp. 66-71). IEEE.

Hiemstra, D (2017). Information Retrieval Models. Retrivited 2 decamber 2017 from http://wwwhome.cs.utwente.nl/~hiemstra/papers/IRModels Tutorial-draft.pdf

Kent, A. & Lancou, H (1968). *Encyclopedia of Library and Information Science*. New York: M. Dekker.

Lewandowski, D. (2012). *Web search engine research*: Emerald Group Publishing Limited.

Miranda, S. V. & Tarapanoff, K. M. A. (2007). Information needs and information competencies: a case study of the off-site supervision of financial institutions in Brazil. *Information Research, 13*(2), Retrieved from http://InformationR.net/ir/13-2/paper344.html

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting with Computers, 10*(3), 303-320.

Reitz, J. M. (2019). Information need. In Online Dictionary for Library and Information Science. Retrived from https://www.abc-clio.com/ODLIS/odlis_i.aspx.

Saracevic, T. (2007). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology , 58* (13), 2126-2144.

Saracevic, T. (2012). Research on relevance in information science: a historical perspective. *Proceedings of the ASAS&T2012 on pre-conference on the history of ASAS&T in information and technology.*

Shi, X., & Yang, C. C. (2007). Mining related queries from web search engine query logs using an improved association rule mining model. *Journal of the American Society for Information Science and Technology, 58*(12), 1871-1883.

Siegmund, David O (2019). Probability theory. In *Encyclopedia Britannica. Retrieved* from https://www.britannica.com/science/probability-theory

Thornley, C., & Gibb, F. (2007). A dialectical approach to information retrieval. *Journal of documentation, 63* (5), 755-764.

Timmins, F. (2006). Exploring the concept of 'information need'. I*nternational journal of nursing practice, 12*(6), 375-381

Ullah, M. I. (27 December 2013). *Time Series Analysis*. Basic Statistics and Data Analysis. WEN Themes. Retrieved 2 January 2014

Vidinli, I. B., & Ozcan, R. (2016). New query suggestion framework and algorithms: A case study for an educational search engine. *Information Processing & Management.*

Wilson, T. D. (2000). Recent trends in user studies: action research and qualitative methods. I*nformation research, 5*(3), 5-3.