دانشگاه شهید بهشتی

دانشکده مهندسی و علوم کامپیوتر

راه‌کاری مبتنی بر یادگیری ماشین برای پیش‌بینی مهارت‌های افراد در آینده در وبسایت‌های پرسش‌وپاسخ اجتماعی

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر
گرایش نرم‌افزار

نگارش

سینا محمدی

استاد راهنما

دکتر محمود نشاطی

تابستان ۱۴۰۴

دانشگاه شهید بهشتی

دانشکده مهندسی و علوم کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر - گرایش نرم‌افزار

تحت عنوان:

راه‌کاری مبتنی بر یادگیری ماشین برای پیش‌بینی مهارت‌های افراد در آینده در وبسایت‌های پرسش‌وپاسخ اجتماعی

در تاریخ پایان‌نامه دانشجو، سینا محمدی، توسط کمیته تخصصی داوران مورد بررسی و تصویب نهایی قرار گرفت.

۱- استاد راهنما اول:	نام و نام‌خانوادگی	امضا
۲- استاد راهنما دوم:	نام و نام‌خانوادگی	امضا
۳- استاد مشاور:	نام و نام‌خانوادگی	امضا
۴- استاد داور (داخلی):	نام و نام‌خانوادگی	امضا
۵- استاد داور (خارجی):	نام و نام‌خانوادگی	امضا
۶- نماینده تحصیلات تکمیلی:	نام و نام‌خانوادگی	امضا

کلیه حقوق مادی مترتب بر نتایج مطالعات،
ابتکارات و نوآوری‌های ناشی از تحقیق موضوع
این پایان‌نامه متعلق به دانشگاه شهید بهشتی
می‌باشد.

به نام خدا

نام و نام خانوادگی: سینا محمدی

عنوان پایان‌نامه: راه‌کاری مبتنی بر یادگیری ماشین برای پیش‌بینی مهارت‌های افراد در آینده در وبسایت‌های پرسش‌وپاسخ اجتماعی

استاد راهنما: دکتر محمود نشاطی

اینجانب سینا محمدی تهیه‌کننده پایان‌نامه کارشناسی ارشد حاضر، خود را ملزم به حفظ امانت‌داری و قدردانی از زحمات سایر محققین و نویسندگان بنابر قانون Copyright می‌دانم. بدین وسیله اعلام می‌نمایم که مسئولیت کلیه مطالب درج شده با اینجانب می‌باشد و در صورت استفاده از اشکال، جداول و مطالب سایر منابع، بلافاصله مرجع آن ذکر شده و سایر مطالب از کار تحقیقاتی اینجانب استخراج گشته است و امانت‌داری را به صورت کامل رعایت نموده‌ام. در صورتی که خلاف این مطلب ثابت شود، مسئولیت کلیه عواقب قانونی با شخص اینجانب می‌باشد.

نام و نام خانوادگی: سینا محمدی

تاریخ و امضا:

چکیده

چکیده در اینجا قرار میگیرد

واژگان کلیدی:

فهرست مطالب

۱ کلیات **۱**

۲ ادبیات تحقیق **۲**

۱.۲ مقدمه ۳

۲.۲ بازیابی اطلاعات ۳

۳.۲ بازیابی تخصص ۴

۴.۲ خبره ۵

۵.۲ یادگیری ماشین ۵

۶.۲ یادگیری عمیق ۶

۷.۲ سری‌های زمانی ۶

۸.۲ انجمن‌های پرسش‌وپاسخ ۷

۹.۲ جمع‌بندی ۷

۳ کار های مرتبط **۸**

۱.۳ مقدمه ۹

۲.۳ بازیابی تخصص ۹

۱.۲.۳ یافتن متخصص ۹

۲.۲.۳ نمایه‌سازی ۱۲

۳.۲.۳ مراحل بازیابی تخصص ۱۴
۳.۳ پیش‌بینی سری‌های زمانی ۱۵
۴.۳ جمع‌بندی ۱۷

۴ روش پیشنهادی ۱۸
۱.۴ مقدمه ۱۹
۲.۴ تعریف مسئله ۱۹
۱.۲.۴ روش پایه ۱۹
۲.۲.۴ توضیح و تعاریف ۲۱
۳.۴ ساخت مجموعه داده ۲۲
۱.۳.۴ خوشه بندی ۲۲
۴.۴ پیش‌پردازش و مهندسی ویژگی ۲۴
۱.۴.۴ نرمال‌سازی ۲۶
نرمال‌سازی بر اساس سال فعالیت کاربر ۲۶
نرمال‌سازی بر اساس تمام مهارت‌های کاربر در تمام سال‌ها ۲۶
نرمال‌سازی بر أساس تمامی سال‌ها، تمام کاربران به تفکیک مهارت ۲۶
نرمال‌سازی بر اساس تمامی سال‌ها، تمامی کاربران و تمامی حوزه‌های مهارتی . ۲۷
نرمال‌سازی بر اساس حوزه‌های مهارتی در هر سال، بدون تفکیک کاربر ۲۷
نرمال‌سازی بر اساس داده‌های تمام کاربران و تمام حوزه‌های مهارتی در هر سال ۲۷
۲.۴.۴ تحلیل و دلیل انتخاب روش نهایی ۲۸
۵.۴ جدا سازی داده‌های آزمون از داده‌های آموزش مدل ۲۹
۶.۴ انتخاب مدل ۲۹
۷.۴ آموزش مدل ۳۰
۸.۴ چارچوب آداپتور برای پیش‌بینی بین‌مقیاسی ۳۰

خط‌لولهٔ آداپتور ۳۱
۱.۸.۴ جمع‌بندی ۳۲

۵ نتیجه گیری و کار های آینده ۳۵

مراجع ۳۸

واژه‌نامه انگلیسی به فارسی ۴۲

واژه‌نامه فارسی به انگلیسی ۴۳

فهرست تصاویر

۱.۳ نمایش دهنده تکنیک‌های ساخت نمایه به تفکیک و همچنین روش‌های پیاده سازی هر تکنیک است [۲۸]. ۱۴

۱.۴ نمودار جریان یک کاربر نمونه ۲۰

۲.۴ حوزه‌های مهارتی تولید شده با استفاده از الگوریتم خوشه‌بندی ۲۳

۳.۴ نشان دهنده نمودارهای histogram و plot qq برای حالت های مختلف نرمال‌سازی است. نمودار اول از بالا روش نرمال‌سازی بر اساس سال فعالیت کاربر را نشان می‌دهد. همانطور که مشاهده می‌کنید بیشتر داده‌ها اعداد یک و صفر هستند. نمودار دوم نرمال‌سازی بر اساس تمام مهارت‌های کاربر در تمام سال‌ها است. همانطور که مشخص است تعداد اعداد غیر از صفر و یک کمی بیشتر شده ولی بازهم درصد بزرگی از داده‌ها را این اعداد تشکیل می‌دهند. نمودار دوم و سوم نرمال سازی بر اساس تمام کاربران در تمام سال‌ها را نشان می‌دهند که بیشتر اعداد در این روش نزدیک به صفر هستند و توزیع مناسبی ندارند. نمودار پنجم و ششم نرمال سازی بر اساس سال برای تمام کاربران را نشان می‌دهد. همانطور که مشخص است کمی تعداد اعداد نزدیک به صفر کمتر شده و نمودار اعداد دیگری را نیز نمایش می‌دهد. شکل آخر اعمال مقیاس لگاریتمی روی داده‌ها را نمایش می‌دهد که باعث ایجاد توزیع نرمال در داده‌ها شده و نمودار plot qq بیشتر داده‌ها را روی محور اصلی نمایش می‌دهد. همچنین نمودار histogram به نیز بیانگر این توزیع مناسب است. ۳۳

۴.۴ رشد کاربران stackoverflow ۳۴

فهرست جداول

۱.۴ نام‌های تولید شده به وسیله مدل‌های زبانی بزرگ برای هر حوزه مهارتی خوشه‌بندی شده را نشان می‌دهد. ۲۴
۲.۴ مقایسه روش‌های مختلف نرمال‌سازی و تعداد مقادیر صفر و یک در داده‌ها ۲۸

فصل ۱

کلیات

فصل ۲

ادبیات تحقیق

۱.۲ مقدمه

در این بخش ضمن آشنایی اولیه با مفاهیم کلی و بررسی کلیدواژه‌های استفاده شده در ادامه پایان‌نامه، دید بهتری نسبت به مفاهیم بازیابی اطلاعات، داده‌کاوی و همچنین یادگیری ماشین که در ادامه فصل‌های پایان‌نامه، برای بررسی مسئله مطرح شده در فصل قبل و ارائه راهکار مناسب استفاده می‌شود، کسب خواهید کرد.

۲.۲ بازیابی اطلاعات

بازیابی اطلاعات به طور بنیادی به فرایند یافتن اطلاعات مفید و مرتبط با پرسش کاربر از میان یک مجموعه اسناد اطلاق می‌شود. کاربرد اصلی این فرآیند در موتورهای جست‌وجو، سیستم‌های توصیه‌گر[۱]، تحلیل شبکه‌های اجتماعی و موارد از این دست است. این فرایند به دلیل عوامل متعددی ذاتاً چالش‌برانگیز است؛ از جمله این‌که اطلاعات موجود در پایگاه‌های اسنادی معمولاً ساختارمند نیستند، اسناد در زبان طبیعی و به‌صورت آزاد نوشته می‌شوند و موضوعات بسیار متنوعی را پوشش می‌دهند. در نتیجه، سیستم‌های سنتی بازیابی اطلاعات بیشتر بر شناسایی اسنادی تمرکز داشته‌اند که احتمال می‌رود حاوی اطلاعات مفید باشند، تا ارائه مستقیم پاسخ به پرسش‌های خاص کاربر. روش کلی در این سیستم‌ها شامل نمایش هر سند و پرسش کاربر به‌صورت یک بردار ویژگی (مجموعه‌ای از ویژگی‌ها با وزن‌های مرتبط)، محاسبه شباهت میان آن‌ها، و سپس ارائه اسناد دارای بیشترین شباهت به کاربر است. در گذشته، نمایه‌سازی[۲] به‌صورت دستی و با تخصیص کلیدواژه‌ها انجام می‌شد، اما با افزایش اندازه پایگاه‌های داده این رویکرد ناکارآمد شد و روش‌های خودکار نمایه‌سازی توسعه یافتند. با رشد انفجاری اینترنت و افزایش سریع تعداد کاربران، میزبان‌ها و وب‌سایت‌ها، نیاز به بازیابی اطلاعات کارآمد در وب به‌طور چشمگیری افزایش یافت. بررسی‌ها نشان می‌دهد حدود ۸۵ درصد از کاربران اینترنت برای یافتن اطلاعات خاص از موتورهای جستجو استفاده می‌کنند. با این حال، همین بررسی‌ها نارضایتی قابل توجهی از عملکرد موتورهای جستجوی اولیه گزارش کرده‌اند؛ از جمله سرعت پایین بازیابی، تأخیرهای ارتباطی، و کیفیت پایین نتایج شامل نویز (اطلاعات غیرمرتبط) و پیوندهای معیوب. بازیابی اطلاعات در وب با پیچیدگی‌های جدیدی

---

[۱] system recommendation

[۲] indexing

همراه بود، زیرا حجم اسناد عمومی در وب به‌مراتب از پایگاه‌های داده کلاسیک بیشتر بوده و ماهیت پویای وب نمایه‌سازی جامع و مداوم را به یک چالش بزرگ تبدیل کرده است. برای مقابله با این چالش‌ها، مدل‌ها و تکنیک‌های متنوعی توسعه یافته‌اند. در مجموع، بازیابی اطلاعات نقشی حیاتی در دسترسی به داده‌ها در عصر دیجیتال دارد. با وجود موانع موجود، انتظار می‌رود که با بهره‌گیری از فناوری‌های نوین مانند هوش مصنوعی، این حوزه در آینده شاهد تحولات بیشتری باشد [۱] [۲] [۳].

۳.۲ بازیابی تخصص

بازیابی تخصص، شاخه‌ای از علم اطلاعات است که به شناسایی و رتبه‌بندی افراد بر اساس دانش و مهارت آن‌ها در یک حوزه خاص می‌پردازد. این حوزه به‌ویژه در مدیریت دانش، سیستم‌های توصیه‌گر و تحلیل شبکه‌های اجتماعی اهمیت دارد. هدف اصلی آن، یافتن افرادی است که بتوانند به سؤالات تخصصی پاسخ دهند یا در پروژه‌های پیچیده مشارکت کنند. این فرآیند معمولاً با تحلیل اسناد متنی، مانند مقالات علمی یا پست‌های آنلاین، و گاهی با بررسی روابط حرفه‌ای و اجتماعی افراد انجام می‌شود. برای مثال، در محیط‌های آکادمیک، تعداد استنادها و همکاری‌ها می‌تواند معیاری برای سنجش تخصص باشد. روش‌های بازیابی تخصص به سه دسته اصلی تقسیم می‌شوند: مبتنی بر متن، مبتنی بر شبکه و ترکیبی. روش‌های متنی از تکنیک‌های پردازش زبان طبیعی برای ارزیابی محتوای تولیدشده توسط افراد بهره می‌برند. در مقابل، روش‌های مبتنی بر شبکه بر ساختار روابط، مانند همکاری‌های مشترک یا تعاملات در پلتفرم‌های حرفه‌ای، تمرکز دارند. روش‌های ترکیبی که هر دو جنبه را در نظر می‌گیرند، اغلب دقت بیشتری ارائه می‌دهند. یکی از چالش‌های کلیدی این حوزه، تمایز میان تخصص واقعی و فعالیت‌های صرفاً نمایشی است که ممکن است از حضور پررنگ در شبکه‌های اجتماعی ناشی شود. علاوه بر این، مسائل اخلاقی و حریم خصوصی در استفاده از داده‌های شخصی نیز بحث‌برانگیز است و نیاز به توجه ویژه دارد. کاربردهای بازیابی تخصص شامل پیشنهاد متخصصان در سیستم‌های توصیه‌گر، شناسایی کارشناسان در سازمان‌ها و تقویت همکاری‌ها در جوامع علمی است. به طور کلی، بازیابی تخصص پتانسیل بالایی برای بهبود تصمیم‌گیری و همکاری در محیط‌های مختلف دارد و نیازمند تحقیقات بیشتری است. این بخش از پایان‌نامه تلاش کرده تا دیدگاهی جامع و در عین حال قابل فهم از این حوزه ارائه دهد. امید است که این مرور بتواند راهنمایی

اولیه برای پژوهشگران علاقه‌مند به این موضوع باشد [۴].

۴.۲ خبره

خبره به فردی گفته می‌شود که در یک یا چند حوزه مهارتی مهارت داری دانش عمیق یا تجربه عملی گسترده است. در مسائل بازیابی تخصص یا خبره یابی، خبره به کسی گفته می‌شود که توانایی حل مسائل تخصصی در یک زمینه مشخص را دارا است. با این حال این تعریف دقیقی از خبره نیست و خبره بودن یک فرد می‌تواند یک مسئله کاملا نسبی باشد. ولی در تحقیقات انجام شده معمولا به فردی که دارای تحصیلات رسمی، گواهی‌نامه‌های حرفه‌ای، تجربه کاری و تجربه زیسته در یک حوزه مهارتی است و منابع معتبر داده این تجربیات را تایید می‌کنند، خبره گفته می‌شود [۵].

۵.۲ یادگیری ماشین

یادگیری ماشین با هدف فراهم کردن چارچوبی نظری برای فهم ساختارها و الگوهای پنهان در داده‌ها شکل گرفته است. یادگیری ماشین در ساده‌ترین سطح خود شامل الگوریتم‌هایی است که از روی داده‌های برچسب‌خورده یا بدون برچسب قواعدی استخراج می‌کنند تا بتوانند بر اساس آن پیش‌بینی یا تحلیل انجام دهند. در روش‌های نظارت‌شده، مدل‌ها مانند رگرسیون‌های خطی یا درخت‌های تصمیم، رابطه‌ی ورودی و خروجی را از داده‌های آموزشی می‌آموزند و قابلیت تعمیم به داده‌های جدید را به دست می‌آورند. خوشه‌بندی و سایر روش‌های بدون نظارت، امکان کشف گروه‌های همگن داخل داده‌ها را فراهم می‌کنند و در مسائلی که اطلاعات برچسب‌دار محدود است، ابزار ارزشمندی هستند. مفهوم «ویژگی‌سازی» نقش محوری در بهبود کیفیت مدل‌ها دارد؛ انتخاب و تبدیل درست ویژگی‌ها می‌تواند عملکرد نهایی را تا چند برابر ارتقا دهد. چالش مهم «بیش‌برازش» هنگامی رخ می‌دهد که مدل به جای فراگیری ساختار کلی، جزئیات داده‌های آموزشی را حفظ کند که برای کنترل آن از روش‌هایی مانند اعتبارسنجی متقابل و تنظیم منظم استفاده می‌شود. در سال‌های اخیر، رویکردهای درتجمع مانند رندوم فارست و گرادیان بوستینگ به دلیل پایداری و دقت بالا، مورد توجه ویژه پژوهشگران عملی قرار گرفته‌اند. ظهور شبکه‌های عصبی عمیق با معماری‌های لایه‌لایه، قلمرو جدیدی برای حل مسائلی مثل

تشخیص تصویر و پردازش زبان طبیعی گشوده است. یادگیری انتقالی نیز امکان استفاده از دانش مدل‌های از پیش‌آموزش‌دیده را در مسائل با داده‌های محدود فراهم کرده است و به کاهش زمان آموزش کمک می‌کند. بحث تفسیرپذیری مدل‌ها به‌عنوان یک ضرورت اخلاقی و کاربردی مطرح است تا تصمیمات ماشینی قابل اعتماد و قابل توضیح باشند. در نهایت، تلفیق روش‌های کلاسیک و نوین بهترین مسیر برای غلبه بر محدودیت‌های هر رویکرد است و در ادامه فصل‌های پژوهش به بررسی جزئیات این ترکیب‌ها و روش‌های پیاده‌سازی عملی خواهیم پرداخت.

۶.۲ یادگیری عمیق

یادگیری عمیق شاخه‌ای از یادگیری ماشین است که براساس تعریف ارائه شده در کتب مرجع، با استفاده از شبکه‌های عصبی چندلایه سعی می‌کند الگوهای پیچیده و غیریکنواخت در داده‌ها را کشف کند. در این رویکرد، هر لایه از شبکه مسئول استخراج ویژگی‌هایی با سطح انتزاعی متفاوت است و ترکیب این ویژگی‌ها به مدل توانایی فهم روابط غیرخطی را می‌دهد. این روش انعطاف پذیری بالاتری نسبت به رویکردهای پایه یادگیری ماشین دارد و در نتیجه امکان پیاده‌سازی راه‌حل برای مسائل پیچیده‌تر و با مشخصه‌های بالاتر را دارا است [۶] [۷] [۸]. لازم به ذکر است که، در این پایان‌نامه نیز از یک شبکه عصبی چندلایه برای حل مسئله پیش‌بینی تخصص استفاده شده است.

۷.۲ سری‌های زمانی

سری‌های زمانی به عنوان چارچوبی برای تحلیل تغییرات کمّی در طول بازه‌های پیوسته زمانی شناخته می‌شوند. در بررسی این داده‌ها، توجه به ساختار درونی، همچون روندهای بلندمدت و نوسانات فصلی، اهمیت ویژه‌ای دارد. ماهیت فرایندهای زمان‌مند می‌تواند ویژگی‌های پویایی و خودهمبستگی پیچیده‌ای را نشان دهد. جداسازی عناصر مختلف یک سری زمانی زمینه را برای شناخت عوامل تأثیرگذار فراهم می‌آورد. داده‌های سری زمانی در حوزه‌های گوناگون علمی، از اقتصاد تا علوم محیطی، کاربرد گسترده‌ای یافته‌اند. بررسی بقای ساختاری و ثبات واریانس در بازه‌های زمانی متفاوت، از جمله چالش‌های تحلیل این سری‌هاست. عامل زمانی به‌عنوان

بُعدی پایه‌ای، فرصت و پیچیدگی تحقیقات تجربی را هم‌زمان افزایش می‌دهد. مطالعات مربوط به تغییر رفتار متغیرها در طول زمان، دیدگاهی دینامیک و واقع‌بینانه ارائه می‌کند. درک توالی‌های زمانی، بنیانی برای تحلیل علّی و تبیین روابط متغیرها به‌حساب می‌آید. تحلیل دقیق سری‌های زمانی دیدی منسجم از سفر داده در طول زمان در اختیار پژوهشگر می‌گذارد. [۹]

۸.۲ انجمن‌های پرسش‌وپاسخ

انجمن‌های پرسش‌وپاسخ به‌عنوان سکویی برای تبادل دانش میان کاربران، امکان پیگیری پرسش‌ها و ارائه پاسخ‌های ساختاریافته را فراهم می‌آورند. در این فضا، هر کاربر می‌تواند هم‌زمان در نقش پرسشگر و پاسخ‌دهنده ظاهر شود که همین تنوع نقش‌ها، به پویایی شبکه کمک می‌کند. محبوبیت چشمگیر پلتفرم‌هایی مانند Stack-Overflow موجب شده پژوهشگران با رویکردی میان‌رشته‌ای به تحلیل محتوایی و ساختاری این سامانه‌ها بپردازند. همچنین کاربران می‌توانند به پاسخ و پرسش دیگر کاربران امتیاز داده و مشخص کنند که این پرسش‌وپاسخ تا چه میزان مفید بوده است. در حوزه بازیابی تخصص یا خبره‌یابی نیز به وفور از داده‌های استخراج شده از این انجمن‌ها برای تحلیل، پژوهش و آزمایش راهکارهای مختلف استفاده می‌شود. [۱۰]

۹.۲ جمع‌بندی

در این فصل با معرفی چندین اصطلاح کاربردی برای ادامه مطالعه پایان‌نامه، یک دید کلی به خواننده در موضوع تحقیقاتی خبره‌یابی یا بازیابی تخصص داده شد. در فصل آینده برای تکمیل دید کلی خواننده نسبت به موضوع، به بررسی پژرهش‌های انجام شده در این حوزه پرداخته خواهد شد.

فصل ۳

کار های مرتبط

۱.۳ مقدمه

در این بخش بیشتر با مفاهیم بازیابی تخصص یا خبره‌یابی آشنا می‌شوید و شناخت نسبی از تحقیقات انجام شده در این حوزه به دست می‌اورید . از طرفی مسئله این پایان‌نامه در واقع یک مسئله پیش‌بینی سری‌های زمانی [۱]است، پس اخرین فعالیت‌های انجام شده در این حوزه با تمرکز بر کاربرد یادگیری‌ماشین در پیش‌بینی سری‌های زمانی بررسی خواهد شد.

۲.۳ بازیابی تخصص

به طور کلی حوزه بازیابی تخصص یا خبره‌یابی را می‌توان به دو دسته بزرگ تقسیم کرد، یافتن متخصص[۲] و نمایه‌سازی از تخصص‌های یک متخصص[۳] . در بخش یافتن متخصص پژوهشگران تلاش می‌کنند به این سوال پاسخ‌دهند که، «**چه کسانی در موضوع x مهارت دارند؟**» هدف کلی این حوزه یافتن متخصص‌ها در مهارت خاص و امتیازدهی به آن‌ها است. از طرف دیگر در دسته نمایه‌سازی تلاش می‌شود که به سوال «**فرد x دارای چه مهارت‌هایی است؟**» پاسخ دهند. و هدف کلی این دسته یافتن میزان تخصص یک فرد در حوزه‌های مهارتی گوناگون و ساخت نمایه از کاربران است [۴] . مسئله مطرح شده در این پایان‌نامه بیشتر زیرمجموعه نمایه‌سازی قرار می‌گیرد. اما از نتایج آن می‌توان در مسائل یافتن متخصص نیز استفاده کرد.

۱.۲.۳ یافتن متخصص

راه‌کارهایی که برای مسئله یافتن متخصص مطرح شده را می‌توان به چهار دسته تقسیم کرد. مدل‌های برپایه تجزیه ماتریس، مدل‌های درختی، مدل‌های برپایه یادگیری عمیق و مدل‌های رتبه‌بندی [۱۱] [۱۲].

مدل‌های مبتنی بر تجزیه ماتریس یکی از دسته‌های اصلی روش‌های مورد استفاده در سامانه‌های توصیه‌گر و به‌طور خاص در مسئله‌ی یافتن کارشناسان در انجمن‌های پرسش‌وپاسخ به‌شمار می‌روند. از دیدگاه ریاضی، تجزیه ماتریس به معنای تجزیه‌ی یک ماتریس داده‌شده -که معمولاً تعامل میان کارشناسان و پرسش‌ها را نمایش

[۱] timeseries
[۲] finding Expert
[۳] profiling Expert

می‌دهد - به حاصل‌ضرب چند ماتریس ساده‌تر و با ابعاد کمتر است. هدف از این فرایند، کشف ویژگی‌های پنهان است که زیربنای روابط میان موجودیت‌هایی مانند کارشناسان و پرسش‌ها را تشکیل می‌دهند. در این چارچوب، هر کارشناس و هر پرسش به فضایی مشترک از عوامل پنهان نگاشته می‌شوند و تعامل میان آن‌ها به‌صورت ضرب داخلی در این فضا مدل‌سازی می‌شود. این مدل‌ها با هدف کمینه‌سازی تابع خطای مربعات، سعی در افزایش دقت یافتن متخصص دارند [؟] [۱۳] [۱۴] [۱۵].

مدل‌های درخت تقویتی نیز به‌عنوان یکی از روش‌های یادگیری ماشینی برای یافتن متخصص در سامانه‌های پرسش‌وپاسخ مطرح شده‌اند. در این روش‌ها، پیش‌بینی نهایی با جمع‌کردن تدریجی چند درخت تصمیم به‌دست می‌آید مدل‌های مبتنی بر یادگیری عمیق، دسته‌ای برجسته از تکنیک‌ها را در فرایند یافتن متخصص در سامانه‌های پرسش‌وپاسخ اجتماعی تشکیل می‌دهند؛ این مدل‌ها با یادگیری بازنمایی‌های با ابعاد بالا از اطلاعات متخصصان، اطلاعات موضوع و تعاملات میان آن دو، عملکرد چشمگیری در سامانه‌های پیشنهاددهی و وظایف تطبیق نشان داده‌اند [۱۲].

روش رتبه‌بندی تلاش می‌کند با استفاده از راهکارهای گوناگون از جمله روش‌های آماری یا روش‌های مبتنی بر یادگیری ماشین و با رتبه‌بندی کاندیداها بهترین افراد را برای یک مهارت خاص پیدا کند [۱۱]. از کارهای انجام شده در این حوزه می‌توان به مقاله نشاطی و همکاران اشاره کرد که با استفاده از الگوریتم svm سعی در رتبه‌بندی کاندیداهای یک مهارت دارد [۱۶]. همچنین در مقاله‌ای به قلم رستمی و همکاران تلاش شده که با تمرکز بر یافتن کارآموز برای مهارت‌های خاص روش‌های نوینی در حوزه یافتن تخصص معرفی کند [۱۷].

در پژرهش‌های دیگر تلاش شده به جای یافتن متخصص در حوزه مهارتی خاص، با رتبه‌بندی افراد، فرد مناسب برای حضور در یک تیم چابک را شناسایی و معرفی کند [۱۸] [۱۹] . همچنین در پژوهش‌های دیگری که تمرکز آن‌ها به جای یافتن یک فرد متخصص، تلاش برای سازمان‌دهی و جمع‌آوری یک تیم از خبره‌ها بوده است، تلاش شده به جای یافتن یک متخصص یا خبره یک گروه حداقلی از متخصصین شناسایی شود که توانایی انجام یک کار چندجانبه را دارا هستند [۲۰].

در حوزه‌ی تشکیل تیم‌های نرم‌افزاری چابک، پژوهش رستمی به مسئله‌ی شناسایی و انتخاب کارشناسان خبره، به‌ویژه کارشناسان t-shaped که دارای دانش عمیق در یک حوزه‌ی تخصصی و دانش عمومی در چند حوزه‌ی

مرتبط هستند، پرداخته است. نویسندگان این پژوهش محدودیت‌های روش‌های موجود را شناسایی کرده‌اند؛ از جمله تکیه بر مدل‌های غیر یادگیری ماشین مبتنی بر قواعد اکتشافی و استفاده‌ی محدود از صرفاً تعداد اسناد کاربران به‌جای تحلیل محتوای آن‌ها برای برآورد دانش t-shaped . برای رفع این مشکلات، رویکردی چندمرحله‌ای ارائه شده است که در گام نخست با بهره‌گیری از مدل پیشرفته‌ی RDM ، فهرست اولیه‌ی نامزدهای خبره انتخاب می‌شود، سپس با استفاده از یک مدل یادگیری عمیق نوآورانه که محتوای پست‌های کاربران در StackOverflow را تحلیل می‌کند، میزان دانش مرتبط آن‌ها مجدداً برآورد می‌شود و در نهایت، اعضای بهینه‌ی تیم با استفاده از مدل برنامه‌ریزی خطی صحیح انتخاب می‌گردند [۲۱] .

در مقاله مروری که در سال ۲۰۲۵ منتشر شده دسته‌بندی جدیدی برای راه‌حل‌های معرفی شده در مسئله یافتن متخصص معرفی شده، که این دسته‌بندی‌ها شامل روش‌های مبتنی بر محتوا ، روش‌های مبتنی بر شبکه و روش‌های ترکیبی است. در روش‌های مبتنی بر متن تمرکز روی یافتن متخصص از روی متن تولید شده در وبسایت‌های پرسش‌وپاسخ است. همچنین در روش‌های مبتنی بر شبکه تمرکز روی یافتن متخصص از طریق ارتباط آن فرد با متخصص‌های از پیش شناخته شده است، و در روش ترکیبی سعی شده که با ادغام دو روش فوق به دقت بالاتری در مسئله یافتن متخصص دست یافت [۲۲].

از کارهایی که در زمینهٔ استفاده از یادگیری ماشین به‌خصوص یادگیری عمیق در زمینهٔ پیش‌بینی تخصص حال حاضر افراد انجام شده، می‌توان به مقالهٔ Flambeau Florentin Kameni Jiechieu، اشاره کرد. تمرکز این مقاله بر روی تکنیک‌های یادگیری ماشین به جهت پیش‌بینی میزان تخصص افراد بر اساس رزومهٔ آن‌هاست. در واقع در این مقاله بخش‌هایی از تخصص‌های فرد که در رزومه به‌صورت مستقیم به آن اشاره نشده است، پیش‌بینی می‌شود و سایر بخش‌های نامفهموم تخصص افراد را کشف می‌کند [۲۳].

در حوزه‌ی یافتن متخصص، یکی از چالش‌های اساسی شناسایی کاربران خبره‌ای است که دانش و مهارت لازم برای پاسخ‌گویی به پرسش‌های خاص را داشته باشند؛ زیرا تعداد این افراد معمولاً کمتر از تعداد پرسش‌گران است. برای حل این مسئله‌ی مهم در یافتن کارشناسان و مسیریابی پرسش، قاسمی چارچوبی پیشنهاد کرده‌اند که هم‌زمان از شباهت‌های واژگانی و معنایی بین پرسش‌های جدید و پاسخ‌های موجود، به همراه روابط اجتماعی میان کاربران استفاده می‌کند. این مدل بر پایه‌ی رویکرد مبتنی بر سند که پیش‌تر در یافتن کارشناسان نتایج

امیدوارکننده‌ای نشان داده است، بنا شده و یک معیار شباهت کاربر نیز به آن افزوده شده است تا نامزدها بر اساس حوزه‌های علاقه‌مندی رتبه‌بندی شوند. روش کار شامل ایجاد یک گراف همگن از کاربران سامانه بر اساس گراف چهاروجهی شامل پرسش‌گر، پرسش، پاسخ و پاسخ‌دهنده است که در آن گره‌ها کاربران و یال‌ها زمانی ایجاد می‌شوند که یک کاربر به پرسش کاربر دیگری پاسخ دهد [۲۴].

همچنین روش‌های دیگری برپایه ساخت یک شبکه همکاری از افراد خبره و یافتن خبره مناسب بر اساس همکاری‌های انجام شده توسط آن خبره با دیگر خبره‌ها معرفی شده و روشی معرفی شده که متخصصان در حیطه کاری مرتبط به هم شناسایی شوند. این روش با استفاده از داده‌های همکاری‌های انجام شده در فعالیت‌های پژوهشی (مثلا همکاری چند پژوهشگر در انتشار یک مقاله با موضوع خاص) سعی در طراحی و تولید این شبکه همکاری دارد [۲۵].

از دیگر کارهایی که در حوزه یافتن متخصص پژوهشگر انجام شده می‌توان به مقاله almuhana و yafooz اشاره کرد. در این مقاله با استفاده از منابعی مانند مقالات علمی، استنادات، و اطلاعات نویسندگان تلاش کرده‌اند معیارهایی کمی و کیفی برای سنجش تخصص پژوهشگران ارائه دهند. رویکردهای مختلفی همچون روش‌های مبتنی بر محتوا، تحلیل شبکه‌های اجتماعی علمی و مدل‌های یادگیری ماشین برای رتبه‌بندی و شناسایی متخصصان به کار گرفته شده است [۲۶].

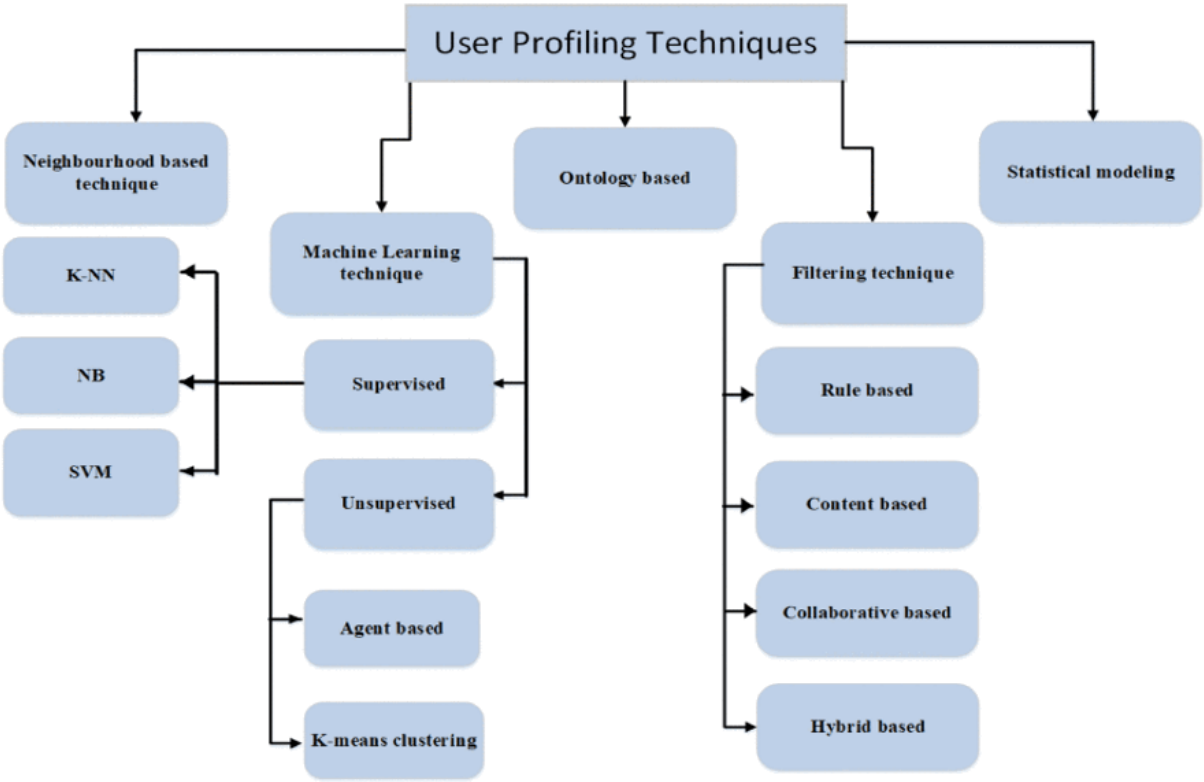
۲.۲.۳ نمایه‌سازی

هدف کلی کارهای این حوزه، پاسخ دادن به این سوال است که یک متخصص چه مهارت‌هایی داشته و دارد. معمولا این سوال پس از یافتن متخصص و نیاز به بررسی دقیق‌تر تخصص‌های آن فرد پرسیده می‌شود [۲۷]. یک مقاله مروری که در حوزه نمایه‌سازی به دسته‌بندی روش‌های ساخت نمایه می‌پردازد، روش‌های ساخت نمایه کاربران را به پنج دسته کلی تقسیم می‌کند. این روش‌ها شامل روش‌های آماری، روش‌های مبتنی بر یادگیری ماشین، روش‌های مبتنی بر فیلتر کردن اطلاعات کاربر ، تکنیک‌های مبتنی بر یافتن همسایگی و روش‌های هستی‌شناسانه است [۲۸]. روش‌های آماری با استفاده از کلمات کلیدی یا لاگ‌های کاربر، یک نمایه کاربری ایجاد می‌کند. در سامانه‌های وب، این کار ممکن است شامل کلمات بسیار پرتکرار در صفحات وب بازدیدشده باشد. برای مثال، یک مدل می‌تواند بر اساس توییت‌ها، نشانی‌های اینترنتی مورد علاقه، افراد دنبال‌شده و

رأی‌دهی اجتماعی در میان همسایگی کاربر، آدرس‌های اینترنتی را به کاربران توییتر پیشنهاد دهد. همچنین می‌تواند الگوهای استفاده را از لاگ‌های سامانه‌های رایانه‌ای استخراج کند تا ناهنجاری‌ها را شناسایی نماید، هرچند این امر دامنه تحلیل را به رفتارهای کاربری ثبت‌شده محدود می‌کند [۲۹]. در روش فیلترینگ با معرفی یک سری قوانین، اطلاعات موجود برای یک کاربر به گونه‌ای فیلتر می‌شود که اطلاعات مورد نیاز باقی‌مانده و اطلاعات غیرمرتبط با کاربر حذف می‌شود، در نهایت یک نمایه باکیفیت از کاربر باقی می‌ماند [۳۰]. استفاده از یادگیری ماشین در نمایه‌سازی در پژرهش‌های متفاوتی معرفی شده، اما به طور کلی در این روش‌ها مدلی معرفی می‌شود که به عنوان ورودی اطلاعات بدون برچسب شامل اطلاعات رزومه، لایک‌ها و دیسلایک‌های دریافت شده در شبکه‌های اجتماعی و غیره دریافت شده و نمایه کاربر به عنوان خروجی مدل معرفی می‌شود [۲۸]. تکنیک مبتنی بر همسایگی رویکردی است که از مفهوم «دنبال کردن دوستان» یا گروه‌هایی از دوستان با علایق مشترک (که به آن همسایگی گفته می‌شود) استفاده می‌کند. هدف اصلی این روش، ایجاد یک فهرست مرتب از کاربران برای هر کاربر است که اعضای آن از نظر علایق و ویژگی‌ها به کاربر شباهت دارند. این روش می‌تواند با رفع کمبود اطلاعات در نمایش علایق شخصی، به ساخت نمایه کاربر کمک کند و معمولاً برای این کار از تاریخچه مرور کاربر استفاده می‌شود. به عنوان نمونه، از این روش برای تخمین موقعیت مکانی کاربران توییتر در سطح شهر با تحلیل ویژگی‌ها، همبستگی‌ها و روابط برچسب‌خورده در یک شبکه اجتماعی استفاده شده است [۳۱].

در روش مبتنی بر آنتولوژی برای پروفایل‌سازی کاربر، از آنتولوژی‌ها به‌عنوان یک مفهوم‌سازی از حوزه مورد نظر در قالبی قابل فهم برای انسان و قابل پردازش توسط ماشین استفاده می‌شود که شامل موجودیت‌ها، ویژگی‌ها، روابط و قواعد است. این رویکرد به‌ویژه با پیشرفت وب معنایی[۱] ، پتانسیل بالایی برای بهبود پروفایل‌سازی کاربر دارد. اهمیت این روش در آن است که چالش‌های مرتبط با اشتراک‌گذاری علایق کاربر - که یکی از مشکلات رایج در نمایش پروفایل کاربر است - را برطرف می‌کند و امکان نمایش مؤثر علایق کاربر و تبادل اطلاعات مرتبط در میان سامانه‌های مختلف را فراهم می‌آورد [۳۲].

[۱] وب معنایی یا web sementic مفهومی است که توسط تیم برنرز-لی، خالق وب جهان‌گستر، مطرح شد و هدف آن ارتقاء وب موجود به محیطی است که در آن داده‌ها دارای معنا و ساختار مشخص برای ماشین‌ها باشند. در این رویکرد، اطلاعات به‌گونه‌ای توصیف می‌شوند که نه‌تنها برای انسان، بلکه برای نرم‌افزارها نیز قابل درک، پردازش و استنتاج باشند.



شکل ۱.۳: نمایش دهنده تکنیک‌های ساخت نمایه به تفکیک و همچنین روش‌های پیاده سازی هر تکنیک است [۲۸].

از دیگر کارهای انجام شده در زمینه یافتن پروفایل، می‌توان به مقاله دهقان اشاره کرد که تلاش می‌کند با استخراج نمایه افراد t-shaped به سازمان‌ها و شرکت‌ها برای جذب بهتر نیروی انسانی کمک کند [۳۳]. در مطالعه‌ای توسط سیلوا و همکاران رویکردی مبتنی بر گراف برای شناسایی نمایه افراد ارائه شده است. هدف اصلی این پژوهش، استخراج حوزه‌های تخصصی افراد بدون نیاز به برچسب‌گذاری قبلی بوده است؛ این موضوع به‌ویژه برای شناسایی گروه‌های پژوهشی در یک سازمان و بررسی شباهت میان حوزه‌های تخصصی پژوهشگران اهمیت دارد . زیرا با این روش می‌توان با اطلاعات جمع‌اوری شده از منابع مختلف نمایه کامل‌تری از کاربران تولید کرد [۳۴]. در دیگر کارهای انجام شده، روند تغییرات نمایه افراد در طول زمان ساخته‌شده و این تغییرات با استفاده از نمودارهای مختلف به تصویر کشیده می‌شود [۳۵] [۳۶] .

۳.۲.۳ مراحل بازیابی تخصص

طبق دسته‌بندی دیگری که بر اساس مقاله مروری منتشر شده به قلم Gonçalves و Dorneles است، مراحل بازیابی تخصص را می‌توان به چهار دسته کلی تقسیم کرد. این چهار دسته شامل شناسایی منابع داده، استخراج

داده ، دسته‌بندی و پیش‌پردازش داده‌های استخراج شده و تحلیل داده‌های استخراج است [۳۷].

۳.۳ پیش‌بینی سری‌های زمانی

داده‌های سری زمانی، که به‌عنوان مجموعه‌ای از مشاهدات زمانی تعریف می‌شوند، یکی از حوزه‌های برجسته پژوهش در یادگیری ماشین و هوش مصنوعی هستند. این داده‌ها نقش حیاتی در تفسیر پدیده‌های پیچیده دنیای واقعی و شناسایی روندهای آینده ایفا می‌کنند. این مجموعه داده‌ها در حوزه‌های گوناگونی از جمله مدل‌سازی اقلیم، علوم زیستی، پزشکی، خرده‌فروشی و مالی و به‌طورکلی هر حوزه‌ای که بعد زمان در آن مطرح است، به‌طور گسترده‌ای وجود دارند؛ جایی که پیش‌بینی دقیق می‌تواند پایه‌ای برای کاربردهایی مانند برنامه‌ریزی تولید و بهینه‌سازی فراهم کند .تکامل روش‌های پیش‌بینی سری‌های زمانی از رویکردهای آماری کلاسیک به سمت تکنیک‌های یادگیری عمیق پیش رفته است که ناشی از افزایش حجم داده‌ها، ابعاد بالا و پیچیدگی آن‌هاست .از آن‌جایی که مسئله مطرح شده در این پایان‌نامه به صورت تئوری یک مسئله پیش‌بینی سری‌های زمانی است، لازم است پژوهش‌های انجام شده در این حوزه بررسی شود تا راه‌حل ارائه شده منطبق بر آخرین پیشرفت‌ها در این زمینه باشد.

از پژوهش‌های انجام شده در این زمینه می‌توان به مقالهٔ Liu و همکاران اشاره کرد که در آن روش‌های سنتی سری‌های زمانی مانند مدل خودرگرسیو، میانگین متحرک و مدل تلفیقی به‌عنوان پایه‌ای‌ترین تکنیک‌ها معرفی شده است. در ادامه، با توسعه مدل ARIMA [۱] ، فصل جدیدی در پیش‌بینی داده‌های غیرایستا گشوده شد تا بتوان نوسانات بلندمدت را دقیق‌تر مدلسازی کرد. با ظهور داده‌های حجیم و پیچیده، محققان به سمت روش‌های

[۱] مدل ARIMA یا خودرگرسیونی یکپارچه با میانگین متحرک یکی از پرکاربردترین مدل‌های آماری در تحلیل سری‌های زمانی است. این مدل ترکیبی از سه مؤلفه اصلی دارد:

- خودرگرسیونی که از مقادیر گذشته متغیر برای پیش‌بینی مقدار فعلی استفاده می‌کند،
- میانگین متحرک که خطاهای پیشین مدل را در نظر می‌گیرد،
- تفاضل‌گیری که برای ایستا کردن سری زمانی به کار می‌رود.

یادگیری ماشین مثل شبکه‌های عصبی مصنوعی و ماشین بردار پشتیبان [۱] سوق یافتند. سپس، استفاده از شبکه‌های بازگشتی به ویژه شبکه‌های حافظه بلندمدت[۲] ، امکان یادگیری وابستگی‌های طولانی‌مدت را فراهم ساخت [۳۸].

از دیگرکارهایی که جدای از تکنیک‌های آماری، بر جمع‌آوری و بررسی روش‌های مبتنی بر یادگیری‌ماشین در پیش‌بینی سری‌های زمانی تمرکز دارند، می‌توان به مقالهٔ محمود و محمد اشاره کرد که روش‌های پارامتریک کلاسیک (از جمله مدل خودرگرسیو ، هموارسازی نمایی و مدل‌های ساختاری سری‌های زمانی) را پایه کار خود قرار داده‌اند. این پژوهش نشان می‌دهد که تکنیک‌های یادگیری عمیق، برخاسته از موفقیت‌هایشان در دسته‌بندی تصاویر، پردازش زبان طبیعی و یادگیری تقویتی، می‌توانند الگوهای پیچیده زمانی را به‌خوبی مدل‌سازی کنند. اخیراً، مدل‌های مبتنی بر مکانیزم توجه با توانایی درک روابط بلندمدت در داده‌های ترتیبی و ترکیب آن‌ها با اجزاء آماری کلاسیک در مدل‌های هیبرید، دقت پیش‌بینی را به‌طور چشمگیری ارتقا داده‌اند [۳۹].

در زمینه پیش‌بینی سری‌های زمانی از طریق یادگیری عمیق نیز می‌توان به مقالهٔ Han و همکاران اشاره کرد که پژوهش‌های پیشین را در حوزه پیش‌بینی سری‌های زمانی مرور و تحلیل می‌کند. در این مقاله، مدل‌های کلاسیک مانند ،ARIMA روش‌های مبتنی بر فیلترینگ و ماشین‌های بردار پشته‌آی [۳] به‌عنوان پایه‌ای‌ترین تکنیک‌ها معرفی شده‌اند. همچنین شبکه‌های عصبی سطح‌پایین شامل پرسپترون‌های چندلایه و شبکه‌های بازگشتی سنتی، به همراه مدل‌های مولد مانند فرآیندهای گاوسی و شبکه‌های بیزی بررسی شده‌اند. این روش‌های قدیمی اغلب در مواجهه با داده‌های حجیم و پیچیده ناتوان بودند و محدودیت‌هایی در نمایش کارآمد توابع پیچیده داشتند. مقاله

[۱] ماشین بردار پشتیبان یا SVM یکی از الگوریتم‌های قدرتمند یادگیری نظارت‌شده است که برای مسائل طبقه‌بندی و رگرسیون به کار می‌رود.

[۲] شبکه‌های حافظه بلندمدت کوتاه‌مدت یکی از گونه‌های پیشرفته شبکه‌های عصبی بازگشتی هستند که برای مدل‌سازی داده‌های دنباله‌دار مانند سری‌های زمانی، متن و گفتار طراحی شده‌اند. برخلاف شبکه‌های عصبی معمولی که با مشکل ناپدید شدن گرادیان مواجه می‌شوند و نمی‌توانند وابستگی‌های بلندمدت را به‌خوبی حفظ کنند، شبکه‌های حافظه بلندمدت با استفاده از ساختاری خاص به‌نام سلول حافظه و سه دروازه (ورودی، خروجی و فراموشی) قادر است اطلاعات مهم را برای بازه‌های زمانی طولانی حفظ کرده و اطلاعات غیرضروری را حذف کند. این ویژگی باعث شده شبکه‌های حافظه بلندمدت در بسیاری از کاربردهای پیش‌بینی سری‌های زمانی، ترجمه ماشینی و تحلیل احساسات عملکرد بسیار موفقی داشته باشد.

[۳] ماشین‌های بردار پشته‌ای نوعی معماری ترکیبی هستند که از چندین مدل ماشین بردار پشتیبان به‌صورت لایه‌ای یا سلسله‌مراتبی استفاده می‌کنند.

حاضر با به‌کارگیری تکنیک‌های یادگیری عمیق، این خلأها را پر کرده و ظرفیت مدل‌سازی و پیش‌بینی الگوهای چندبعدی و بزرگ را به‌طور چشمگیری ارتقا می‌دهد [۴۰].

۴.۳ جمع‌بندی

در این فصل ضمن آشنایی با حوزه تحقیقاتی بازیابی تخصص یا خبره‌یابی به دو زیربخش یافتن متخصص و نمایه‌سازی پرداخته شد. همچنین بررسی روی تحقیقات انجام شده در حوزه سری‌های زمانی و یادگیری عمیق در این حوزه انجام شد. در فصل آینده روش پیشنهادی برای مسئله پیش‌بینی تخصص معرفی خواهد شد.

فصل ۴

روش پیشنهادی

۱.۴ مقدمه

در این فصل ابتدا در بخش ۴-۲ به بیان مسئله و فرضیات پیرامون آن پرداخته می‌شود. سپس در بخش ۴-۳ و ۴-۲ مراحل جمع‌آوری داده و پیش‌پردازش روی آن معرفی می‌شود. و در بخش ۴-۵ مدلی برای پیش‌بینی مهارت‌های افراد معرفی می‌شود و در نهایت مراحل و نتایج آموزش مدل در بخش ۴-۶ جمع آوری می‌شود.

۲.۴ تعریف مسئله

در مسئله پیش‌بینی تخصص ورودی مسئله روند تغییرات مهارت‌های یک فرد در یک بازه زمانی (مثلا در این پایان‌نامه از باز ۲۰۰۸ تا ۲۰۱۴ میلادی استفاده شده است) و خروجی آن میزان خبرگی آن فرد در سال آینده (در این پایان نامه از سال ۲۰۱۵ میلادی استفاده شده است) است. در واقع در این مسئله نمایه کاربر در یک بازه زمانی چند ساله به عنوان ورودی به مدل معرفی شده داده شده و سپس نمایه آن کاربر در سال جدید توسط مدل پیش‌بینی می‌شود. در شکل ۱.۴ یک کاربر نمونه و میزان تغییرات تخصص‌های این فرد در بازه زمانی توسط یک نمودار جریان نمایش داده شده. هدف اصلی این مسئله پیش‌بینی سال جدید با استفاده از داده‌های سال‌های گذشته است به طوری که بیشتری شباهت را با تخصص‌های واقعی افراد داشته باشد.

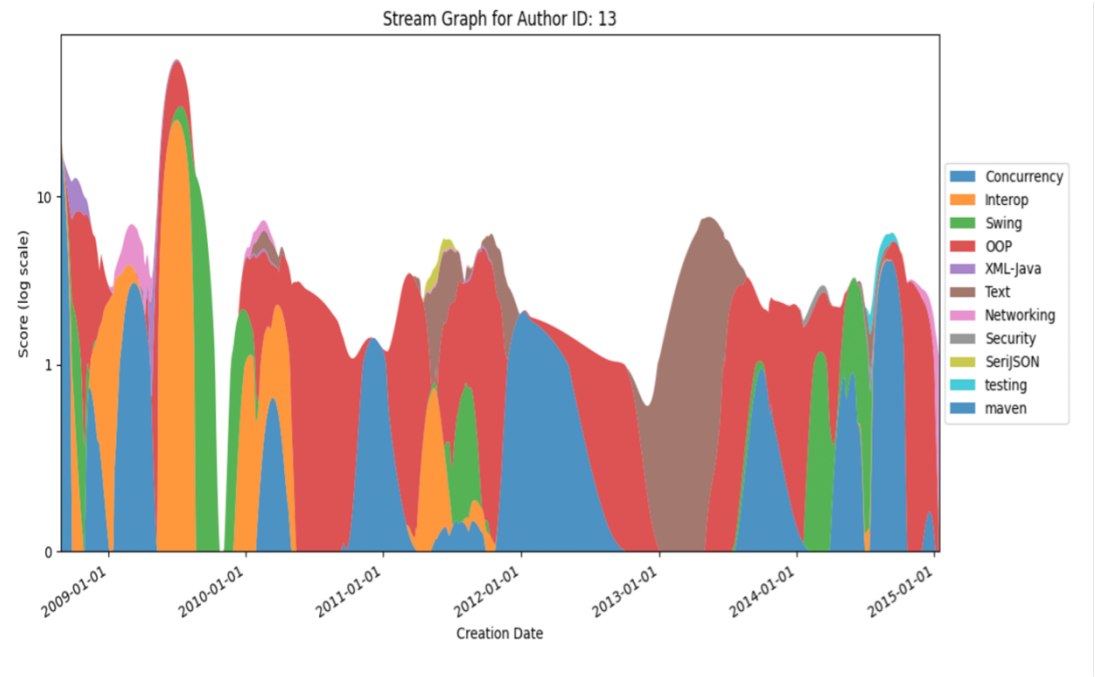
۱.۲.۴ روش پایه

در بسیاری از مسائل پیش‌بینی، انتخاب و تعریف یک روش پایه اهمیت زیادی دارد، زیرا امکان مقایسه و ارزیابی عملکرد روش‌های پیشنهادی را فراهم می‌سازد. در واقع، روش پایه به عنوان یک معیار حداقلی عمل می‌کند که نشان می‌دهد آیا روش پیشنهادی قادر است عملکردی بهتر از یک مدل ساده و ابتدایی ارائه دهد یا خیر.

با توجه به اینکه در زمینه مورد مطالعه این پایان‌نامه تاکنون پژوهش مستقیمی انجام نشده است، در تحقیقات پیشین روش پایه مشخصی معرفی نشده است. از این رو، لازم است با بهره‌گیری از تکنیک‌های آماری ساده، روش پایه‌ای تعریف شود تا نتایج مدل‌های پیشرفته‌تر قابل مقایسه باشند.

برای تعریف روش پایه، رویکردهای متنوعی قابل تصور هستند؛ به عنوان مثال:

- در نظر گرفتن مقدار سال جدید دقیقاً برابر با مقدار سال قبل؛



شکل ۱.۴: در این نمودار هر رنگ نماینده یک حوزه مهارتی است و محور عمودی میزان تخصص افراد بر اساس معیار score (میزان امتیازی که کاربران دیگر به پاسخ این کاربر در سایت stackoverflow داده اند) را نمایش می‌دهد. همچنین محور افقی سال‌های فعالیت این کاربر بین ۲۰۰۸ تا ۲۰۱۵ را نمایش می‌دهد.

- برآورد مقدار سال جدید بر اساس میانگین تمام سال‌های گذشته؛
- استفاده از میانگین وزن‌دار که در آن سال‌های اخیر وزن بیشتری نسبت به سال‌های قدیمی‌تر دارند؛
- به‌کارگیری روش‌های پیشرفته‌تر همچون مدل‌های سری زمانی آماری مانند ARIMA به‌عنوان روش پایه.

با وجود امکان استفاده از روش‌های پیچیده‌تر، در این پایان‌نامه به منظور سادگی و ایجاد معیاری شفاف برای مقایسه، روش پایه بر اساس **میانگین مقادیر سال‌های گذشته** انتخاب شده است. در این روش، مقدار پیش‌بینی‌شده برای سال جدید برابر با میانگین مقادیر سال‌های گذشته در نظر گرفته می‌شود. این رویکرد از یک‌سو پیاده‌سازی ساده‌ای دارد و از سوی دیگر معیاری منطقی و قابل اتکا برای ارزیابی عملکرد روش پیشنهادی ارائه می‌کند.

۲.۲.۴ توضیح و تعاریف

تعاریف متغیرها:

u شناسهٔ کاربر .(user)

s شاخص یا نماد یک حوزهٔ مهارتی .(skill)

t, y, y^* شاخص زمان یا سال؛ y^* سالِ پیش‌بینی‌شده (target year).

a یک عنصر فعالیت (مثلاً یک پاسخ یا پست) مربوط به کاربر در یک مهارت و سال.

$A_{u,s,t}$ مجموعهٔ تمام عناصر a مربوط به کاربر u در مهارت s و سال t.

$\text{score}(a)$ مقدار امتیازِ اختصاص‌یافته به عنصر a (مثلاً تعداد رأی‌های مثبت).

$x_{u,s,t}$ مقدار تجمعیِ خامِ امتیازها برای کاربر u، مهارت s و سال t (مثلاً $x_{u,s,t} = \sum_{a \in A_{u,s,t}} \text{score}(a)$).

$\tilde{x}_{u,s,t}$ مقدار تبدیل‌شدهٔ لگاریتمیِ $x_{u,s,t}$ (مثلاً $\tilde{x} = \log(1+x)$).

$a_{u,s,y}$ نماد دیگری برای مقدار خامِ تجمعیِ کاربر-مهارت-سال (معادلِ $x_{u,s,t}$ در بخش‌های دیگر متن).

$T_{s,y}$ مجموعِ مقادیرِ یک مهارت s در سال y بر روی همهٔ کاربران: $T_{s,y} = \sum_u a_{u,s,y}$.

$U_{u,y}$ مجموعِ مقادیرِ همهٔ مهارت‌ها برای کاربر u در سال y: $U_{u,y} = \sum_s a_{u,s,y}$.

ε ثابتِ کوچک برای پایداری عددی و جلوگیری از تقسیم بر صفر (مثلاً 10^{-8}).

$n^{(6)}_{u,s,y}$ مقدار نُرم‌شده طبق استراتژی ۶ (Skill-Year Normalization): $n^{(6)}_{u,s,y} = \dfrac{a_{u,s,y}}{T_{s,y} + \varepsilon}$.

$n^{(1)}_{u,s,y}$ مقدار نُرم‌شده طبق استراتژی ۱ (User-Year Normalization): $n^{(1)}_{u,s,y} = \dfrac{a_{u,s,y}}{U_{u,y} + \varepsilon}$.

$\widehat{n}^{(6)}_{u,s,y^*}$, $\widehat{\mathbf{n}}^{(6)}_{u,y^*}$ بردار/مؤلفه‌های پیش‌بینی‌شده در فضای نُرم‌شدهٔ Strategy ۶.

$\widehat{a}_{u,s,y^*}$ پیش‌بینی بازگردانده‌شده به مقیاس خام برای کاربر u، مهارت s و سال y^*.

$\widehat{\mathbf{n}}^{(1)}_{u,y^*}$, $\widehat{n}^{(1)}_{u,s,y^*}$ پیش‌بینیِ نُرم‌شدهٔ نهایی در فضای Strategy ۱ برای ارزیابی.

$\mathbf{n}^{(1)}_{u,y^*}$ بردار پروفایل واقعیِ کاربر u در سال y^* پس از نُرم‌سازی Strategy ۱.

$\|\cdot\|_2$ نُرمِ اقلیدسی (Euclidean norm) بردار: $\|\mathbf{v}\|_2 = \sqrt{\sum_i v_i^2}$.

$\mathrm{cosine}(\cdot,\cdot)$ کسینوس شباهت بین دو بردار؛ معیارِ ارزیابی برای تطابقِ توزیع مهارت‌ها.

۳.۴ ساخت مجموعه داده

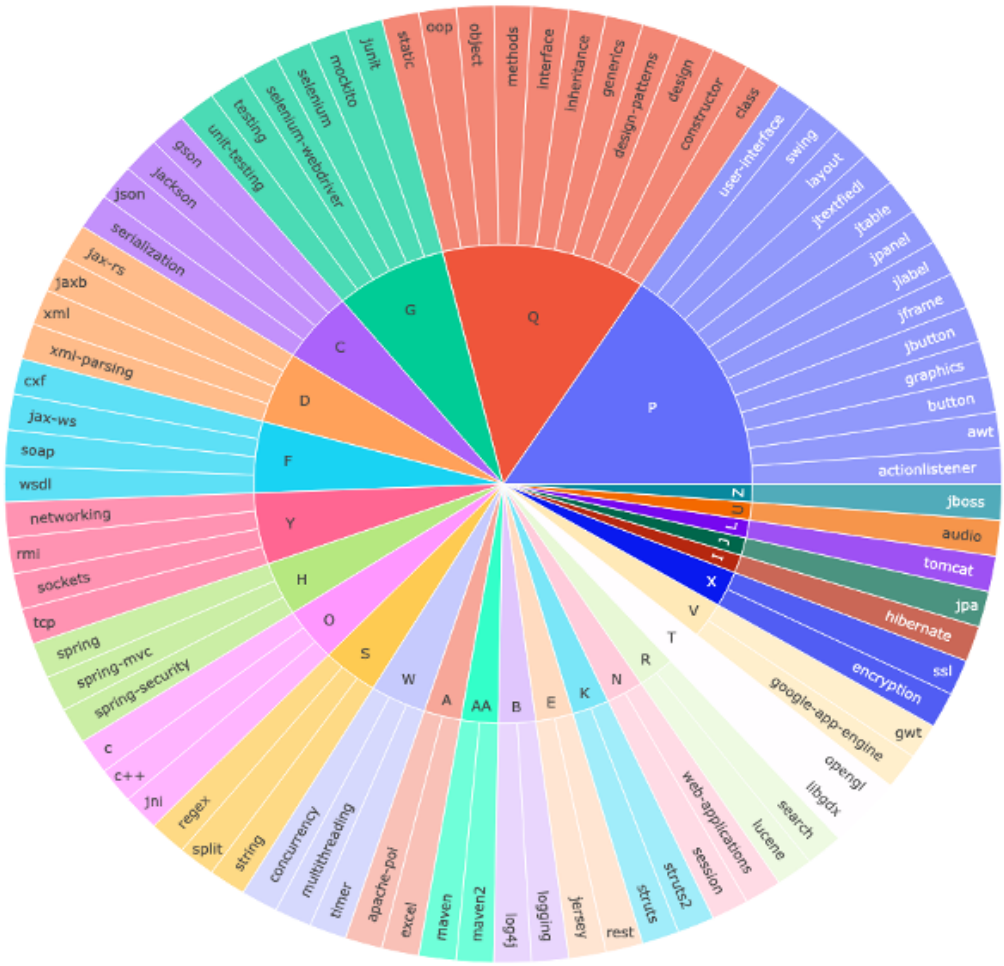
همانطور که پیشتر گفته شد برای این پایان‌نامه از داده‌های خام وبسایت stackoverflow استفاده شده است. داده‌های مورد استفاده در این پژوهش از وب‌سایت Stack استخراجOverflow شده است. این داده‌ها شامل متن پرسش‌ها، متن پاسخ‌ها، امتیاز هر پرسش و پاسخ، تاریخ ثبت پرسش یا پاسخ، برچسب‌های مرتبط با هر پرسش (نمایانگر موضوع یا حوزه‌ی مطرح‌شده) و همچنین شناسه‌ی کاربری ایجادکننده‌ی پرسش یا پاسخ می‌باشد. با این حال، برای ساخت مجموعه‌داده‌ی مناسب به منظور حل مسئله‌ی پیش‌بینی تخصص کاربران، تنها به چهار ویژگی کلیدی نیاز بود: شناسه‌ی کاربر، حوزه‌ی مهارتی، تاریخ فعالیت و امتیاز. تمام این اطلاعات (به جز حوزه‌های مهارتی) در مجموعه‌داده‌ی اولیه وجود داشت. برای استخراج حوزه‌های مهارتی از کاربران، مطابق با رویکرد ارائه‌شده در مقاله‌ی رستمی و همکاران، پرسش‌ها و پاسخ‌ها با استفاده از یک الگوریتم خوشه‌بندی گروه‌بندی شدند تا حوزه‌های مهارتی حاصل گردد [۱۷].

۱.۳.۴ خوشه بندی

برای انجام خوشه‌بندی، پرسش‌ها و پاسخ‌های مرتبط با زبان برنامه‌نویسی Java انتخاب شدند. دلیل این انتخاب، حجم بالاتر داده‌های مربوط به Java نسبت به سایر زبان‌ها و فناوری‌ها در StackOverflow بود. فرآیند خوشه‌بندی با استفاده از روش خوشه‌بندی تجمیعی[۱] انجام گرفت و در نتیجه برچسب‌های مشابه در یک خوشه قرارگرفتند. برچسب‌های مرتبط با هر حوزه‌ی مهارتی در شکل ۲.۴ ارائه شده‌اند. با توجه به اینکه الگوریتم‌های

[۱] Clustering Agglomerative

خوشه‌بندی معمولاً نام‌های قابل درکی برای انسان تولید نمی‌کنند، در این پژوهش از مدل‌های زبانی[۱] بزرگ برای نام‌گذاری خوشه‌ها استفاده شده تا حوزه‌های مهارتی نام‌گذاری بهتری را داشته باشند. نام‌های نهایی خوشه‌ها در جدول ۱.۴ نمایش داده شده‌اند.



شکل ۲.۴: در این شکل تمام خوشه‌های تولید شده با الگوریتم خوشه‌یابی به همراه مهارت‌های مرتبط با هر خوشه نمایش داده شده است.

[۱] LLMs

Name	SkillArea	Name	SkillArea	Name	SkillArea
Graphics	T	jpa	J	ExcelPOI	A
audio	U	struts	K	Log۴J	B
Cloud	V	tomcat	L	SeriJSON	C
Concurrency	W	WebSessions	N	XML-Java	D
Security	X	Interop	O	RESTful	E
Networking	Y	Swing	P	SOAP-based	F
jboss	Z	OOP	Q	testing	G
maven	AA	Search	R	spring	H
		Text	S	hibernate	I

جدول ۱.۴: نام‌های تولید شده به وسیله مدل‌های زبانی بزرگ برای هر حوزه مهارتی خوشه‌بندی شده را نشان می‌دهد.

۴.۴ پیش‌پردازش و مهندسی ویژگی

در حال حاضر داده‌های موجود شامل چهار ویژگی هستند: شناسه کاربر، تاریخ فعالیت، امتیاز کاربر و حوزه مهارتی. تاریخ فعالیت کاربر یک تاریخ دقیق شامل سال، ماه، روز و ساعت فعالیت است؛ در نتیجه یک کاربر ممکن است در یک سال بیش از یک داده داشته باشد. از آنجا که در مسئلهٔ مطرح‌شده قصد داریم مهارت کاربر را به ازای هر سال فعالیت بررسی کرده و در نهایت مهارت‌های کاربر را در سال بعد پیش‌بینی کنیم، لازم است که تمام امتیازهای یک کاربر در یک حوزهٔ مهارتی و در یک سال تجمیع شود.

برای تجمیع امتیاز کاربران، شش تابع وجود دارد که می‌تواند در تجمیع این ویژگی به ما کمک کند:

- **تابع تعداد:** این تابع تعداد سطرهای داده در یک سال فعالیت کاربر را محاسبه می‌کند.
- **تابع جمع:** این تابع امتیازهای کاربر را در یک سال و در یک حوزهٔ مهارتی خاص جمع می‌کند.
- **تابع میانگین:** این تابع میانگین امتیاز کاربر را در یک سال و در یک حوزهٔ مهارتی محاسبه می‌کند.

- **تابع میانه:** این تابع میانهٔ امتیاز کاربر را در یک سال و در یک حوزهٔ مهارتی محاسبه می‌کند.
- **تابع بیشینه:** این تابع بیشینهٔ امتیاز کاربر را در یک سال و در یک حوزهٔ مهارتی محاسبه می‌کند.
- **تابع کمینه:** این تابع کمینهٔ امتیاز کاربر را در یک سال و در یک حوزهٔ مهارتی محاسبه می‌کند.

در بررسی‌های انجام‌شده، دو تابعِ **میانگین** و **جمع** نتایج مناسب‌تری برای این مسئله نشان دادند و برای این پژوهش از تابع تجمیعی **جمع** استفاده شده است. دلیل این انتخاب این است که علاوه بر میزان امتیاز دریافتی هر کاربر در سال، تعداد دفعاتی که آن کاربر در هر حوزهٔ مهارتی تعامل داشته است نیز دارای اهمیت است؛ و تابع تجمیعی جمع می‌تواند توازنی میان دفعات تعامل و مجموع امتیازهای کاربر برقرار کند، در حالی که تابع میانگین وزنِ دفعات تعامل را در نظر نمی‌گیرد.

$$x_{u,s,t} = \sum_{a \in A_{u,s,t}} \text{score}(a)$$

در نهایت مجموعهٔ داده شامل: سال فعالیت، مجموعهٔ امتیاز کاربر (تجمیع‌شده)، شناسهٔ کاربر و حوزهٔ مهارتی او است. همان‌طور که مشخص است، برای آموزش بیشتر مدل‌های مبتنی بر یادگیری ماشین نیاز به نرمال‌سازی داده‌های عددی است. دادهٔ عددی موجود در این مسئله، عدد تجمیع‌شدهٔ امتیاز کاربران است. برای نرمال‌سازی از تکنیک **min-max** استفاده شده است.

اما مسئلهٔ اصلی روش نرمال‌سازی است. برای ساخت پروفایل کاربر در این مسئله از روشی استفاده شده که مجموع مهارت‌های کاربر در تمام حوزه‌ها در یک سال نسبت به هم نرمال می‌شوند؛ در نتیجه با مشاهدهٔ پروفایل کاربر در هر سال می‌توان سهم هر حوزهٔ مهارتی در نمایهٔ کاربر در آن سال را به صورت یک عدد بین صفر و یک بررسی کرد. بنابراین برای ساخت پروفایل کاربر، این روش نرمال‌سازی گزینهٔ مناسبی به نظر می‌رسد.

با این حال، در تلاش‌های انجام‌شده برای آموزش مدل با استفاده از این روش نرمال‌سازی، یک مشکل بزرگ مشاهده شد: وجود مقادیر بیش از حدِ برابر با 0 یا 1 مطلق. این مشکل باعث بروز مسائل در فرآیند آموزش مدل از جمله عدم بهبود مدل در هر دوره و در نتیجه کاهش دقت مدل می‌شود. بنابراین نیاز دیده شد که روش‌های دیگر نرمال‌سازی نیز بررسی و آزمایش شوند.

۱.۴.۴ نرمال‌سازی

در این پژوهش، روش‌های نرمال‌سازی داده‌ها به پنج دسته‌ی کلی تقسیم شدند که هر یک مزایا و محدودیت‌های خاص خود را دارند. فرض کنید $x_{u,s,t}$ مقدار خام حوزهٔ مهارتی s برای کاربر u در سال t باشد. ما شش روش نرمال‌سازی را بررسی می‌کنیم:

نرمال‌سازی بر اساس سال فعالیت کاربر

در این روش، امتیازهای هر کاربر در یک سال نسبت به یکدیگر نرمال می‌شوند. مزیت اصلی این رویکرد کاهش اثرات تجمع داده‌ها و نوسانات امتیاز در طول یک سال است و تا حدی مشکل ایستایی در یادگیری را کاهش می‌دهد. با این حال، این روش به تنهایی قادر به حل کامل مشکل نیست و برخی داده‌ها همچنان در بازه‌های اکستریم (صفر یا یک مطلق) باقی می‌مانند.

$$\hat{x}_{u,s,t} = \frac{x_{u,s,t}}{\sum_{s' \in S} x_{u,s',t}}$$

نرمال‌سازی بر اساس تمام مهارت‌های کاربر در تمام سال‌ها

در این رویکرد، امتیازهای یک کاربر در تمامی سال‌های فعالیت نسبت به یکدیگر مقیاس‌بندی می‌شوند. این روش باعث می‌شود تغییرات مهارتی کاربر در طول زمان قابل مقایسه‌تر شوند و مشکل ایستایی در یادگیری تا حدی کاهش یابد. با این حال، تأثیر تغییرات جمعیتی کاربران و رشد کلی پلتفرم بر امتیازات ثبت‌شده همچنان بر جا می‌ماند و نمی‌توان آن را به‌طور کامل خنثی کرد.

$$\hat{x}_{u,s,t} = \frac{x_{u,s,t}}{\sum_{s' \in S} \sum_{\tau \in T} x_{u,s',\tau}}$$

نرمال‌سازی بر أساس تمامی سال‌ها، تمام کاربران به تفکیک مهارت

در این روش، هر حوزه مهارتی یک کمینه (عدد ۰) و یک بیشینه (عدد ۱) دریافت کرده و باقی سطرهای مرتبط با این مهارت عدد بین صفر و یک دریافت می‌کنند. در این روش بیشتر داده‌های تولید شده نزدیک به عدد صفر

هستند.

$$\hat{x}_{u,s,t} = \frac{x_{u,s,t}}{\sum_{u' \in U} \sum_{s' \in S} \sum_{\tau \in T} x_{u',s',\tau}}$$

نرمال‌سازی بر اساس تمامی سال‌ها، تمامی کاربران و تمامی حوزه‌های مهارتی

در این روش، بیشترین امتیاز موجود در کل مجموعه‌داده به مقدار ۱ و کمترین امتیاز به مقدار صفر نگاشت می‌شوند و سایر مقادیر بین این دو مقدار قرار می‌گیرند. مزیت این رویکرد، مقیاس‌بندی یکپارچه تمام داده‌ها است. اما مشکل اصلی، توزیع نامناسب داده‌ها است؛ به گونه‌ای که اکثر داده‌ها در بازه‌ی ۰ تا ۲۵.۰ قرار می‌گیرند و تقریباً هیچ داده‌ای بین ۷۵.۰ تا ۱ مشاهده نمی‌شود. این نابرابری در توزیع باعث می‌شود مدل آموزش داده‌شده در بازه‌های بالاتر داده‌ها ضعیف عمل کند.

$$\hat{x}_{u,s,t} = \frac{x_{u,s,t}}{\sum_{u' \in U} \sum_{\tau \in T} x_{u',s,\tau}}$$

نرمال‌سازی بر اساس حوزه‌های مهارتی در هر سال، بدون تفکیک کاربر

در این رویکرد، داده‌ها برای هر حوزه‌ی مهارتی در هر سال مستقل نرمال می‌شوند، بدون آنکه هویت کاربران در نظر گرفته شود. این روش توزیع مناسب‌تری از داده‌ها ایجاد می‌کند و اثرات ناشی از تغییر تعداد کاربران در طول زمان را کاهش می‌دهد.

$$\hat{x}_{u,s,t} = \frac{x_{u,s,t}}{\sum_{u' \in U} \sum_{s' \in S} x_{u',s',t}}$$

نرمال‌سازی بر اساس داده‌های تمام کاربران و تمام حوزه‌های مهارتی در هر سال

این روش مشابه روش چهارم است، با این تفاوت که تمام کاربران و حوزه‌های مهارتی در مقیاس‌بندی هر سال لحاظ می‌شوند. نتیجه این نرمال‌سازی، ایجاد توزیع نرمال‌تر و یکنواخت‌تر داده‌ها و کاهش اثرات انحراف ناشی از داده‌های بیش‌نمایش است.

$$\hat{x}_{u,s,t} = \frac{x_{u,s,t}}{\sum_{u' \in U} x_{u',s,t}}$$

روش نرمال‌سازی	تعداد صفرها	تعداد یک‌ها	نسبت صفرها به کل	نسبت یک‌ها به کل
نرمال‌سازی بر اساس سال فعالیت کاربر	۲۰۰۶۴۷	۴۹۴۸۹	۶۲۸۲.۰	۱۵۵.۰
سازی بر اساس تمام مهارت‌های کاربر در تمام سال‌ها	۱۶۶۴۳۴	۴۳۶۴۳	۵۲۱۱.۰	۱۳۶۷.۰
سال‌ها، تمامی کاربران و به تفکیک حوزه‌های مهارتی	۴۰	۲۶	۰۰۰۱.۰	۰۰۰۱.۰
امی سال‌ها، تمامی کاربران و تمام حوزه‌های مهارتی	۱	۱	۰	۰
سال‌سازی بر اساس تمام حوزه‌های مهارتی در هر سال	۱۵	۸	۰	۰
س حوزه‌های مهارتی در هر سال، بدون تفکیک کاربر	۶۰۰	۲۰۹	۰۰۱۹.۰	۰۰۰۷.۰
رتی در هر سال، بدون تفکیک کاربر با مقیاس لگاریتم	۶۰۰	۲۰۹	۰۰۱۹.۰	۰۰۰۷.۰

جدول ۲.۴: مقایسه روش‌های مختلف نرمال‌سازی و تعداد مقادیر صفر و یک در داده‌ها

۲.۴.۴ تحلیل و دلیل انتخاب روش نهایی

روش‌های چهارم و پنجم به لحاظ توزیع نرمال داده‌ها عملکرد بهتری نسبت به سه روش دیگر داشتند. علاوه بر این، دلیل منطقی دیگری نیز وجود دارد که چرا این دو روش مناسب‌تر هستند. وب‌سایت Overflow Stack در طول زمان شاهد رشد قابل توجه کاربران بوده است [۴۱]، [۴۲]؛ این رشد بر تعاملات کاربران تأثیر می‌گذارد و باعث افزایش امتیازات ثبت‌شده برای پرسش‌ها و پاسخ‌های مشابه می‌شود. به بیان دیگر، با افزایش کاربران، ثبت امتیازها برای مهارت‌های مشابه در سال‌های بعدی به طور طبیعی افزایش می‌یابد.

با نرمال‌سازی بر اساس سال، اثر رشد کاربران در طول زمان تا حد زیادی تعدیل می‌شود و مدل آموزش داده‌شده قادر خواهد بود مهارت واقعی کاربران در هر سال را بهتر درک کند، بدون آنکه تحت تأثیر تغییرات جمعیتی و افزایش تعاملات کلی قرار گیرد. به این ترتیب، سهم هر حوزه‌ی مهارتی در پروفایل سالانه‌ی کاربر با دقت بیشتری نمایان می‌شود و مشکل مقادیر صفر یا یک مطلق که در روش‌های اولیه مشاهده شده بود، به طور قابل توجهی کاهش می‌یابد.

همچنین از میان دو روش نهایی روشی که داده را بر اساس سال و به تفکیک هر حوزه مهارتی نرمال‌سازی

می‌کند انتخاب بهتری است. زیرا در وبسایت stackoverflow مهارت‌های پرطرفدار امتیاز بالاتری در پرسش و پاسخ‌ها دریافت می‌کنند. ولی پرطرفدار بودن یک مهارت تاثیری روی میزان تخصص افرادی که روی آن مهارت کار می‌کنند ندارد و در صورت ترکیب تمام حوزه‌های مهارتی در نرمال‌سازی، مدل به اشتباه افرادی که در زمینه مهارت‌های پرطرفدارتر فعالیت داشته‌اند را به عنوان افراد خبره‌تر در نظر می‌گیرد. ولی نرمال سازی به تفکیک حوزه‌های مهارتی اثر این خطا را نیز تعدیل کرده و به مدل در پیش‌بینی بهتر مهارت افراد کمک می‌کند.

۵.۴ جدا سازی داده‌های آزمون از داده‌های آموزش مدل

روش معمول برای جداسازی داده‌های آموزش و آزمون در مسائل یادگیری ماشین تقسیم مجموعه داده به دو بخش معمولا نامساوی، آموزش مدل و سپس ارزیابی مدل با داده‌های آزمون است. این تقسیم معمولا بر اساس سطرهای داده انجام می‌شود. ولی در مجموعه داده موجود از آنجایی که اطلاعات هر کاربر در چندین سطر وجود دارد تقسیم داده‌های تستی بر اساس سطور داده باعث می‌شود که برخی کاربران هم در داده آزمون و هم در داده آموزش حضور داشته باشند. این موضوع باعث می‌شود که نتوان پیش‌بینی دقیق از آن کاربر داشت و مدل نیز اطلاعات ناقصی از کاربرانی که آموزش با آن‌ها انجام می‌شود خواهد داشت. در نتیجه روش بهتر تقسیم بندی کاربران به در گروه ٪۲۰ و ٪۸۰ است. به طوری که کاربران گروه ٪۸۰ برای آموزش مدل استفاده شده و باقی برای ارزیابی و بررسی میزان دقت پیش‌بینی استفاده می‌شوند.

۶.۴ انتخاب مدل

در این پژوهش برای حل مسئله‌ی مورد نظر از یک شبکه‌ی عصبی پیشخور با چند لایه‌ی تمام‌متصل استفاده شد. دلیل انتخاب این ساختار، سادگی، انعطاف‌پذیری و توانایی آن در تقریب توابع غیرخطی پیچیده است. به‌منظور بهبود همگرایی و جلوگیری از بروز مشکلاتی نظیر تغییر توزیع داده‌ها در طول آموزش، از لایه‌های Batch Normalization استفاده شد. همچنین برای کاهش بیش‌برازش در فرآیند آموزش، لایه‌های Dropout به مدل اضافه شدند. این ترکیب به‌عنوان یکی از معماری‌های متداول و کارآمد در مسائل مشابه گزارش شده است و با توجه به محدودیت‌های محاسباتی و داده‌های موجود، انتخاب مناسبی برای این پژوهش محسوب می‌شود.

همچنین لازم به ذکر است برای آموزش مدل از کتاب‌خانه tensorflow و زبان برنامه نویسی پایتون استفاده شده است.

۷.۴ آموزش مدل

برای آموزش مدل از روش‌های اعتبارسنجی متقاطع [۱] استفاده شد تا ارزیابی قابل‌اعتماد و انتخاب مناسب ابرپارامترها ممکن شود؛ تقسیم‌بندی داده‌ها به‌گونه‌ای انجام شد که نمونه‌های مربوط به یک کاربر یا بازهٔ زمانی مشخص در یک بخش قرار گیرند تا از نشت اطلاعات جلوگیری شود. در فرآیند آموزش، معیارهای عددی و برداری مرتبط با هدف نهایی (از جمله معیارهای مربوط به شباهت پروفایل‌ها) پایش شده و تنظیمات مدل بر اساس عملکرد میانگین در تاخورها انتخاب شد.

برای جلوگیری از بیش‌برازش از مکانیزم Stopping Early بهره برده شد؛ مدل در هر تکرار روی مجموعهٔ اعتبارسنجی پایش شده و در صورت عدم بهبود، آموزش متوقف و وزن‌های دورهٔ بهترین عملکرد بازیابی می‌شوند. این ترکیبِ اعتبارسنجی متقاطع و توقف زودهنگام تضمین می‌کند که مدل هم از نظر تعمیم‌پذیری محک زده شود و هم از رشد بیش‌ازحد پیچیدگی در فرایند یادگیری جلوگیری گردد.

۸.۴ چارچوب آداپتور برای پیش‌بینی بین‌مقیاسی

برای همسوسازی روش‌های مختلف نرمال‌سازی که در طی آموزش و ارزیابی به کار می‌روند و از آن‌جایی که ما از دو روش متفاوت نرمال‌سازی برای آموزش مدل و ارزیابی استفاده می‌کنیم، ما یک چارچوب آداپتور توسعه می‌دهیم که به ترتیب (۱) مقادیر خام تجمیع‌شدهٔ مهارت را به فضای ورودی مدل می‌برد (استراتژی ۶)، (۲) پیش‌بینی‌کنندهٔ آموزش‌دیده را در آن فضا اجرا می‌کند تا پیش‌بینی حاصل شود، و (۳) پیش‌بینی را مجدداً به فضای ارزیابی بازمی‌گرداند (استراتژی ۱). این آداپتور تضمین می‌کند که مدل روی نرمال‌سازی‌ای آموزش ببیند که سوگیری ناشی از رشد جمعیت را کاهش می‌دهد (استراتژی ۶)، در حالی که ارزیابی و استفاده‌های بعدی (رتبه‌بندی، شباهت کسینوسی) بر اساس مقیاس نسبی «هر کاربر در هر سال» (استراتژی ۱) انجام می‌شود.

---

[۱] cross-validation

نمادها. فرض کنید $a_{u,s,y}$ مقدار خام تجمیع‌شدهٔ مهارت برای کاربر u، حوزهٔ مهارتی s و سال y باشد (مجموع امتیازهای فعالیت برای آن مثلثِ کاربر-حوزه-سال). مجموع سالانه برحسب هر حوزه را تعریف می‌کنیم:

$$T_{s,y} \;=\; \sum_{u} a_{u,s,y}, \tag{۱.۴}$$

و مجموع سالانهٔ هر کاربر را به‌صورت

$$U_{u,y} \;=\; \sum_{s} a_{u,s,y}. \tag{۲.۴}$$

برای پایداری عددی از ε (ثابت مثبت کوچک، مثلاً 10^{-8}) استفاده می‌کنیم.

نرمال‌سازی استراتژی ۶ (ورودی مدل) به‌صورت نسبت عنصر به عنصر تعریف می‌شود:

$$n^{(6)}_{u,s,y} \;=\; \frac{a_{u,s,y}}{T_{s,y}+\varepsilon}, \tag{۳.۴}$$

و نرمال‌سازی استراتژی ۱ (پروفایل هر کاربر برای ارزیابی) به‌صورت

$$n^{(1)}_{u,s,y} \;=\; \frac{a_{u,s,y}}{U_{u,y}+\varepsilon}. \tag{۴.۴}$$

خط‌لولهٔ آداپتور

برای هر کاربر آزمایشی u و سال پیش‌بینی y^* (مثلاً ۲۰۱۵)، آداپتور مراحل زیر را اجرا می‌کند:

۱. **محاسبهٔ ثابت‌های مقیاس‌دهی.** مقادیر $T_{s,y}$ مربوط به سال‌های مرتبط (به‌ویژه T_{s,y^*}) را بازیابی کنید و بردارهای خام تاریخی کاربر $a_{u,\cdot,y}$ را در اختیار داشته باشید.

۲. **نرمال‌سازی ورودی‌ها (استراتژی ۶).** برای هر سال تاریخی $y < y^*$ که به‌عنوان ورودی قرار می‌گیرد، بردار $\mathbf{n}^{(6)}_{u,y} = \left(n^{(6)}_{u,s,y}\right)_s$ از معادلهٔ (۳.۴) را تشکیل دهید. بردارهای تاریخی را مطابق قالب ورودی مدل، پشته کنید.

۳. **پیش‌بینی در فضای استراتژی ۶.** ورودیِ پشته‌شده را به مدل آموزش‌دیده بدهید و خروجی $\widehat{\mathbf{n}}^{(6)}_{u,y^*} = \left(\widehat{n}^{(6)}_{u,s,y^*}\right)_s$ را دریافت کنید.

۴. **واگرد به مقیاس خام.** پیش‌بینی‌ها را برای سال y^* به مقیاس تجمیع‌شدهٔ خام تبدیل کنید:

$$\widehat{a}_{u,s,y^*} \;=\; \widehat{n}^{(6)}_{u,s,y^*} \cdot \big(T_{s,y^*} + \varepsilon\big).$$

۵. **نرمال‌سازی مجدد با استراتژی ۱.** پروفایل پیش‌بینی‌شدهٔ هر کاربر را برای ارزیابی تهیه کنید:

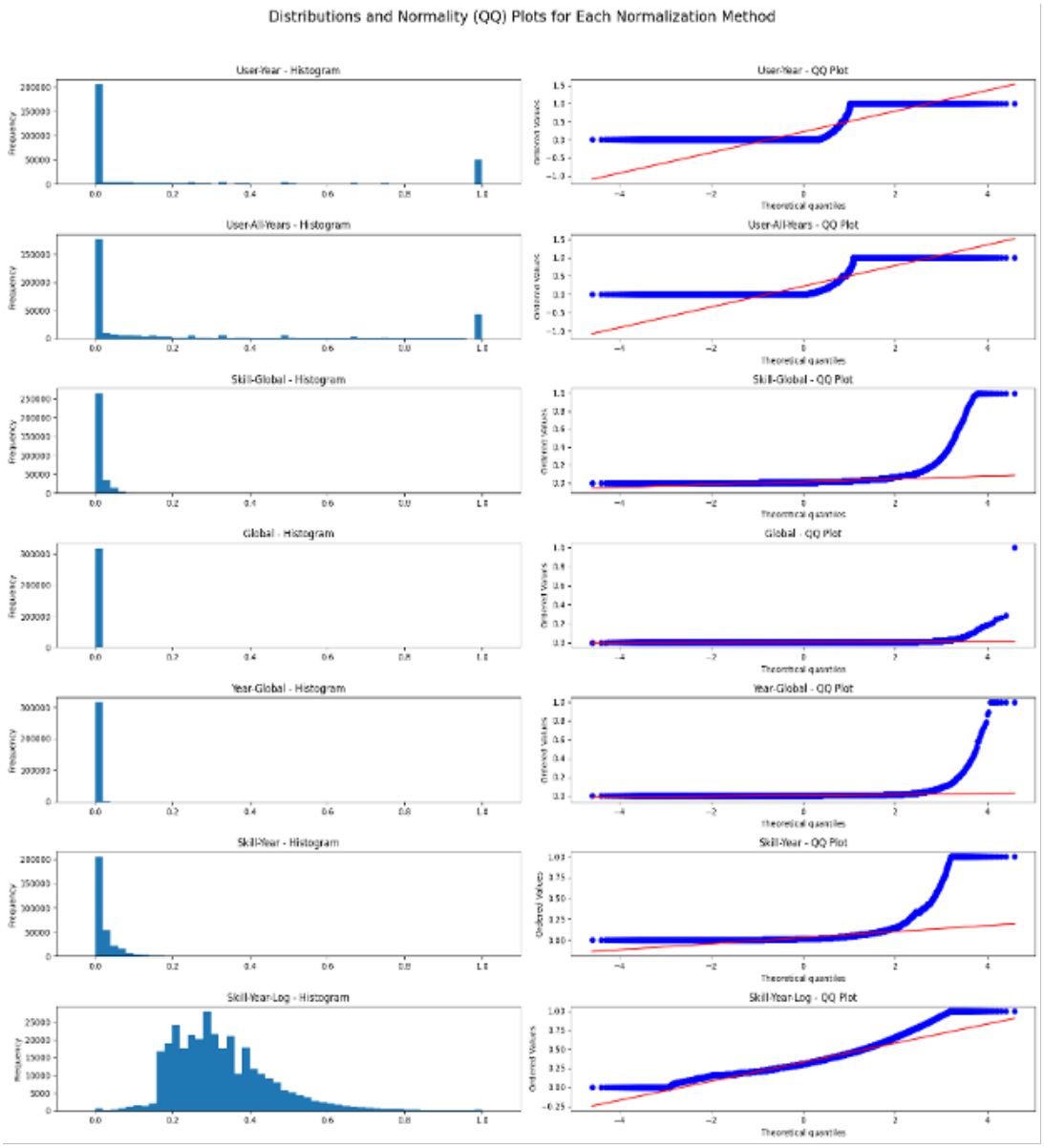
$$\widehat{n}^{(1)}_{u,s,y^*} \;=\; \frac{\widehat{a}_{u,s,y^*}}{\sum_s \widehat{a}_{u,s,y^*} + \varepsilon}.$$

۶. **ارزیابی / پروفایل نهایی.** از $\widehat{\mathbf{n}}^{(1)}_{u,y^*}$ برای محاسبهٔ شباهت کسینوسی نسبت به پروفایل واقعیِ کاربر $\mathbf{n}^{(1)}_{u,y^*}$ استفاده کنید.

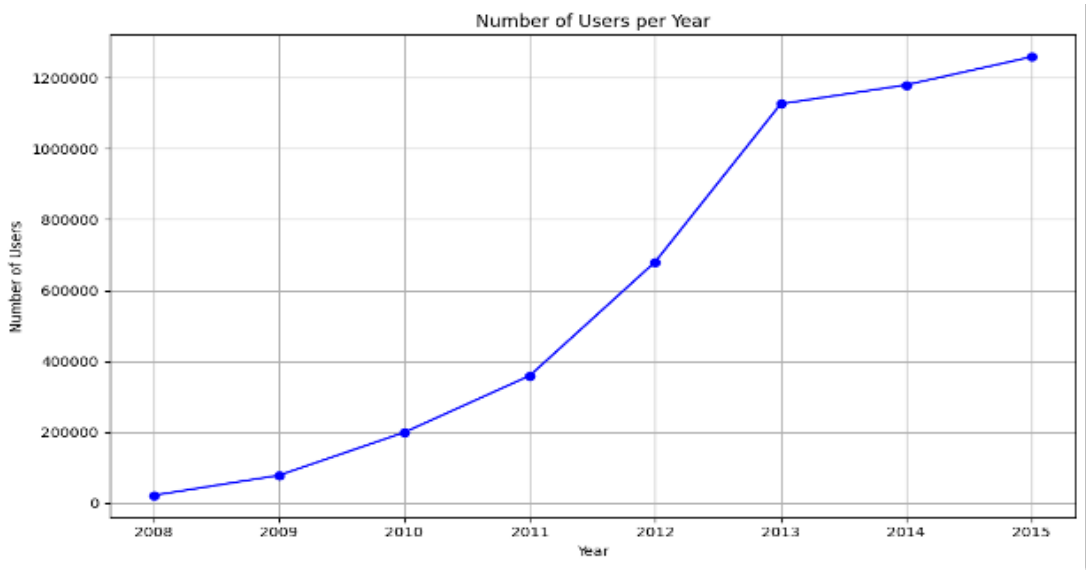
۱.۸.۴ جمع‌بندی

در این فصل روند کلی آماده‌سازی و پردازش داده‌ها برای مسئلهٔ پیش‌بینی بیان شد؛ به‌ویژه روش‌های نرمال‌سازی مورد استفاده برای تبدیل مقادیر خام به فضای ورودی مدل توضیح داده شد. همچنین نوع مدل مورد استفاده و رویکرد آموزشی (شامل تقسیم‌بندی داده‌ها برای اعتبارسنجی و مکانیسم جلوگیری از بیش‌برازش) تشریح شد تا خواننده تصویری جامع از فرایند یادگیری به‌دست آورد. در کنار این‌ها، روش‌های ارزیابی که برای سنجش کیفیت پیش‌بینی‌ها به‌کار رفته‌اند معرفی گردید تا پیوند میان اهداف کاربردی و معیارهای کمّی روشن شود.

در بخش آینده نتایج ارزیابی معرفی شده و این نتایج با روش پایه مقایسه می‌شوند.



شکل ۳.۴: نشان دهنده نمودارهای histogram و plot qq برای حالت های مختلف نرمال‌سازی است. نمودار اول از بالا روش نرمال‌سازی بر اساس سال فعالیت کاربر را نشان می‌دهد. همانطور که مشاهده می‌کنید بیشتر داده‌ها اعداد یک و صفر هستند. نمودار دوم نرمال‌سازی بر اساس تمام مهارت‌های کاربر در تمام سال‌ها است. همانطور که مشخص است تعداد اعداد غیر از صفر و یک کمی بیشتر شده ولی بازهم درصد بزرگی از داده‌ها را این اعداد تشکیل می‌دهند. نمودار دوم و سوم نرمال سازی بر اساس تمام کاربران در تمام سال‌ها را نشان می‌دهند که بیشتر اعداد در این روش نزدیک به صفر هستند و توزیع مناسبی ندارند. نمودار پنجم و ششم نرمال سازی بر اساس سال برای تمام کاربران را نشان می‌دهد. همانطور که مشخص است کمی تعداد اعداد نزدیک به صفر کمتر شده و نمودار اعداد دیگری را نیز نمایش می‌دهد. شکل آخر اعمال مقیاس لگاریتمی روی داده‌ها را نمایش می‌دهد که باعث ایجاد توزیع نرمال در داده‌ها شده و نمودار plot qq بیشتر داده‌ها را روی محور اصلی نمایش می‌دهد. همچنین نمودار histogram به نیز بیانگر این توزیع مناسب است.



شکل ۴.۴: رشد کاربران stackoverflow در گذر زمان بر اساس داده‌های موجود بین سال‌های ۲۰۰۸ تا ۲۰۱۵ میلادی را نمایش می‌دهد. نکته قابل توجه در این شکل این است که دلیل کاهش رشد کاربران در سال ۲۰۱۵ این است که فرآیند جمع آوری داده از این وب‌سایت در اواسط سال ۲۰۱۵ انجام شده و این مجموعه داده شامل تمام داده‌های سال ۲۰۱۵ نیست.

فصل ۵

نتیجه گیری و کار های آینده

ایده و محدوده کاری آینده

در این موضوع با تحلیل داده های خوشه بندی شده ابتدا یک نمایه از هر کاربر که در فرم های اینترنتی مرتبط فعالیت داشته، ساخته میشود. این نمایه توصیف میکند که چند درصد فعالیت کاری هر کاربر به یک موضوع خاص اختصاص داده شده است. سپس یک روش ارزیابی برای امتیازدهی به میزان دقت روش پیشبینی معرفی میشود. به طور کلی سوالی که مطرح میشود و این مقاله سعی در پاسخ دهی به آن دارد این است که بهترین مدل برای پیشبینی تخصص آینده افراد چه مدلی میباشد. این مدل از میان مدل های یادگیری ماشین یا مدل های آماری انتخاب میشود. برای رسیدن به بهترین مدل ابتدا یک روش مناسب برای ارزیابی دقت مدل معرفی میشود. این روش ارزیابی از میان روش های موجود که چند نمونه آن پیشتر مطرح شد انتخاب میشود. سپس چند مدل کاندید انتخاب شده و پیاده سازی آن انجام خواهد شد و نتیجه خروجی آن توسط روش ارزیابی برسی میشود و بهترین مدل یا چند مدل برتر نسبت به باقی مدل ها معرفی میشود. پس میتوان خروجی این مقاله را یک مدل معرفی شده در نظر گرفت.

تمام مدل ها با نمونه ای از داده های سایت stackoverflow آزمون خواهند شد و پیاده سازی هر مدل انجام میشود. در نتیجه میتوان یک مقایسه میان نتیجه پیشبینی مدل ها روی یک داده ی آزمون انجام داد و از آنجایی که این مقایسه در شرایط برابر میان تمام مدل ها انجام میشود، میتوان نتیجه ارزیابی آن را معتبر درنظر گرفت.

این مقایسه از آنجایی دارای اهمیت است که در صورت یافتن بهترین مدل برای پیشبینی این نوع از داده ها در مرحله اول میتوان از آن مدل در کنار کارهایی که پیشتر انجام شده است، روشی معرفی کرد که نتیجه دقیق تری برای جذب افراد در کسب و کار ها استفاده کرد در مرحله دوم از نتیجه داده های تولید شده در این مطالعه برای مطالعه های آینده استفاده شود.

همچنین کارهایی که در آینده میتوان در ادامه این کار انجام داد شامل معرفی روشی که پیاده سازی این مدل را برای صنعت راحت تر کند و همچنین افزودن متغیر های جدید به مدل آموزش دیده شده برای یافتن نتیجه دقیق‌تر میباشد.

نتیجه گیری

به طور کلی مسئله بازیابی نیروی متخصص یک مسئله نسبتا جدید در حوزه بازیابی اطلاعات میباشد. از این رو مسائل باز زیادی همچنان در این حوزه وجود دارد. این مسائل باز بیشتر در حوزه تحلیل داده های استخراج شده، خوشه بندی آن ها و نتیجه گیری از آن‌ها برای یافتن بهترین کاندید یک موقعیت شغلی میباشد.

ضعف اصلی روش های موجود این است که این روش ها معمولا پارامتر های کمی را برای برسی داده های استفاده میکنند و در نتیجه آن خروجی حاصل ممکن است با واقعیت تفاوت داشته باشد. در سال های اخیر سعی شده با معرفی روش هایی دقت خوشه بندی و تحلیل این داده ها افزایش یابد. این مطلب نیز سعی در افزایش دقت روش های پیشین با معرفی روشی جدید برای پیشبینی تخصص آینده افراد دارد تا دقت بازیابی تخصص افراد را افزایش دهد.

کتاب‌نامه

[1] M. Kobayashi and K. Takeda, "Information retrieval on the web," *ACM Comput. Surv.*, vol.32, pp.144-173, June 2000. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[2] M. R. Ghorab, D. Zhou, A. O'Connor, and V. Wade, "Personalised Information Retrieval: survey and classification," *User Modeling and User-Adapted Interaction*, vol.23, pp.381-443, Sept. 2013.

[3] M. Mitra and B. Chaudhuri, "Information Retrieval from Documents: A Survey," *Information Retrieval*, vol.2, pp.141-163, May 2000.

[4] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, "Expertise Retrieval," *Foundations and Trends in Information Retrieval*, vol.6, no.2-3, pp.127-256, 2012.

[5] M. Diaz and A. D. Smith, "What Makes An Expert? Reviewing How ML Researchers Define" Expert"," in *Proceedings of the AAAI| ACM Conference on AI, Ethics, and Society*, vol.7, pp.358-370, 2024.

[6] C. M. Bishop. *Pattern Recognition and Machine Learning* (*Information Science and Statistics*). Berlin, Heidelberg: Springer-Verlag, 2006.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[8] M. Nielsen. *Neural Networks and Deep Learning*. Determination Press, 2015.

[9] G. Kirchg ssner, J. Wolters, and U. Hassler. *Introduction to modern time series analysis*. Springer Science & Business Media, 2012.

[10] M. Neshati, "On early detection of high voted Q&A on Stack Overflow," *Information Processing & Management*, vol.53, pp.780-798, July 2017.

[11] S. Lin, W. Hong, D. Wang, and T. Li, "A survey on expert finding techniques," *Journal of Intelligent Information Systems*, vol.49, pp.255-279, Oct. 2017.

[12] S. Yuan, Y. Zhang, J. Tang, W. Hall, and J. B. Cabota, "Expert finding in community question answering: a review," *Artificial Intelligence Review*, vol.53, pp.843-874, Feb. 2020.

[13] K. Hofmann, K. Balog, T. Bogers, and M. de Rijke, "Contextual factors for finding similar experts," *Journal of the American Society for Information Science and Technology*, vol.61, no.5, pp.994-1014, 2010.

[14] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, (New York, NY, USA), pp.426-434, Association for Computing Machinery, 2008. event-place: Las Vegas, Nevada, USA.

[15] X. Chen and L. Sun, "Bayesian temporal factorization for multidimensional time series prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.44, no.9, pp.4659-4673, 2021.

[16] M. Neshati, Z. Fallahnejad, and H. Beigy, "On dynamicity of expert finding in community question answering," *Information Processing & Management*, vol.53, Sept. 2017.

[17] P. Rostami and M. Neshati, "Intern retrieval from community question answering websites: A new variation of expert finding problem," *Expert Systems with Applications*, vol.181, Nov. 2021.

[18] P. Rostami and M. Neshati, "T-shaped grouping: Expert finding models to agile software teams retrieval," *Expert Systems with Applications*, vol.118, pp.303-308, Mar. 2019.

[19] S. Sotudeh Gharebagh, P. Rostami, and M. Neshati, "T-shaped mining: A novel approach to talent finding for agile software teams," in *40th European Conference on IR Research*, Mar. 2018.

[20] M. Neshati, H. Beigy, and D. Hiemstra, "Expert group formation using facility location analysis," *Information Processing & Management*, vol.50, no.2, pp.361-383, 2014.

[21] P. Rostami and A. Shakery, "A deep learning-based expert finding method to retrieve agile software teams from CQAs," *Inf. Process. Manage.*, vol.60, Mar. 2023. Place: USA Publisher: Pergamon Press, Inc.

[22] M. Azimi, A. Moffat, and J. Zobel, "Expert Finding Revisited: A Uniform Exploration of Methods," in *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '25, (New York, NY, USA), pp.478-487, Association for Computing Machinery, 2025. event-place: Padua, Italy.

[23] K. F. F. Jiechieu and N. Tsopze, "Skills prediction based on multi-label resume classification using CNN with model predictions explanation," *Neural Computing and Applications*, vol.33, pp.5069-5087, May 2021.

[24] N. Ghasemi, R. Fatourechi, and S. Momtazi, "User Embedding for Expert Finding in Community Question Answering," *ACM Trans. Knowl. Discov. Data*, vol.15, Mar. 2021. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[25] S. H. Hashemi, M. Neshati, and H. Beigy, "Expertise retrieval in bibliographic network: a topic dominance learning approach," in *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, (New York, NY, USA), pp.1117-1126, Association for Computing Machinery, 2013. event-place: San Francisco, California, USA.

[26] A. A. Almuhanna and W. M. S. Yafooz, "Expert Finding In Scholarly Data: An Overview," in *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, pp.1-7, 2021.

[27] K. Balog and M. de Rijke, "Determining Expert Profiles (With an Application to Expert Finding)," in *IJCAI*, vol.7, pp.2657-2662, Jan. 2007. Issue: 625.

[28] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, "A Survey of User Profiling: State-of-the-Art, Challenges, and Solutions," *IEEE Access*, vol.7, pp.144907-144924, 2019.

[29] J. Tang, L. Yao, D. Zhang, and J. Zhang, "A Combination Approach to Web User Profiling," *ACM Trans. Knowl. Discov. Data*, vol.5, Dec. 2010. Place: New York, NY, USA Publisher: Association for Computing Machinery.

[30] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: item-to-item collaborative filtering," *IEEE Internet Computing*, vol.7, no.1, pp.76-80, 2003.

[31] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive web search based on user profile constructed without any effort from users," in *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, (New York, NY, USA), pp.675-684, Association for Computing Machinery, 2004. event-place: New York, NY, USA.

[32] L. Han, G. Chen, and M. Li, "A method for the acquisition of ontology-based user profiles," *Adv. Eng. Softw.*, vol.65, pp.132-137, Nov. 2013. Place: GBR Publisher: Elsevier Science Ltd.

[33] M. Dehghan, M. Biabani, and A. A. Abin, "Temporal expert profiling: With an application to T-shaped expert finding," *Information Processing & Management*, vol.56, no.3, pp.1067-1079, 2019. Publisher: Elsevier.

[34] W. Fu and S. Akbar, "Expert Profile Identification From Community Detection on Author-Publication-Keyword Graph With Keyword Extraction," *IEEE Access*, vol.12, pp.27918-27930, 2024.

[35] M. Fazel-Zarandi and M. S. Fox, "Constructing expert profiles over time for skills management and expert finding," in *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, i-KNOW '11, (New York, NY, USA), Association for Computing Machinery, 2011. event-place: Graz, Austria.

[36] M. Iqbal and M. Ahmad, "Ranking and Visualization of Experts for Communication Using Linkedin," in *Proceedings of the Future Technologies Conference (FTC) 2018* (K. Arai, R. Bhatia, and S. Kapoor, eds.), (Cham), pp.1-12, Springer International Publishing, 2019.

[37] R. Gonçalves and C. F. Dorneles, "Automated expertise retrieval: A taxonomy-based survey and open issues," *ACM Computing Surveys (CSUR)*, vol.52, no.5, pp.1-30, 2019.

[38] Z. Liu, Z. Zhu, J. Gao, and C. Xu, "Forecast methods for time series data: A survey," *IEEE Access*, vol.9, pp.91896-91912, 2021.

[39] A. Mahmoud and A. Mohammed, "A survey on deep learning for time-series forecasting," in *Machine Learning and Big Data Analytics Paradigms: Analysis, Applications and Challenges*, pp.365-392, Cham: Springer International Publishing, 2020.

[40] Z. Han, J. Zhao, H. Leung, K. F. Ma, and W. Wang, "A review of deep learning models for time series prediction," *IEEE Sensors Journal*, vol.21, no.6, pp.7833-7848, 2019.

[41] I. Moutidis and H. T. Williams, "Community evolution on stack overflow," *Plos one*, vol.16, no.6, p.e0253010, 2021. Publisher: Public Library of Science San Francisco, CA USA.

[42] G. Livan, G. Pappalardo, and R. Mantegna, "Quantifying the relationship between specialisation and reputation in an online platform," *Scientific Reports*, vol.12, p.16699, 10 2022.

واژه‌نامه انگلیسی به فارسی

کلمه . . . . . Word
کلمه ۲ . . . . . Word 2
کلمه ۳ . . . . . Word 3

واژه‌نامه فارسی به انگلیسی

کلمه ۱ . . . . . word 1
کلمه ۲ . . . . . word 2
کلمه ۳ . . . . . word 3

Abstract:

This is Abstract in English.

Keywords: Word1, Word2, Word3

Shahid Beheshti University
Faculty of Computer Science & Engineering

Semantic Matching Methods on Object Retrieval

By

Arash Dargahi Nobari

A THESIS SUBMITTED
FOR THE DEGREE OF
MASTER OF SCIENCE

Supervisor :

Dr. Mahmood Neshati

2019