



دانشگاه شهید بهشتی

دانشکده مهندسی و علوم کامپیوتر

بررسی روش های پیشبینی روند تخصص افراد از طریق تحلیل داده های سایت Stack overflow

پایان نامه کارشناسی ارشد مهندسی کامپیوتر
گرایش نرم افزار

نگارش

سینا محمدی

استاد راهنما

دکتر محمود نشاطی

تابستان ۱۴۰۳



دانشگاه شهید بهشتی
دانشکده مهندسی و علوم کامپیوتر

پایان نامه کارشناسی ارشد مهندسی کامپیوتر - گرایش نرم افزار
تحت عنوان:

بررسی روش های پیشبینی روند تخصص افراد از طریق تحلیل داده های سایت overflow
Stack

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

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

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 پایان نامه: بررسی روش های پیشبینی روند تخصص افراد از طریق تحلیل داده های سایت overflow

Stack

استاد راهنما: دکتر محمود نشاطی

اینجانب سینا محمدی تهیه کننده پایان نامه کارشناسی ارشد حاضر، خود را ملزم به حفظ امانت داری و قدردانی از زحمات سایر محققین و نویسندگان بنابر قانون Copyright می دانم. بدین وسیله اعلام می نمایم که مسئولیت کلیه مطالب درج شده با اینجانب می باشد 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 امانت داری را به صورت کامل رعایت نموده ام. در صورتی که خلاف این مطلب ثابت شود، مسئولیت کلیه عواقب قانونی با شخص اینجانب می باشد.

نام و نام خانوادگی: سینا محمدی

تاریخ و امضا:

چکیده

بازیابی تخصص یکی از موضوعات پژوهشی در حوزه بازیابی اطلاعات است که به هدف شناسایی نیروی متخصص و تحلیل داده‌های کلان که اغلب به صورت متون غیرساختاریافته هستند، کمک می‌کند. این فرآیند به سازمان‌ها و شرکت‌ها امکان می‌دهد تا نیروی متخصص مورد نیاز خود را به سهولت بیابند. به طور کلی، مباحث مربوط به بازیابی تخصص به چند دسته تقسیم می‌شوند. دسته اول شامل استخراج داده‌ها و آماده‌سازی آن‌هاست که معمولاً این داده‌ها برای اهداف تحقیقاتی از انجمن‌های آنلاین استخراج می‌شوند. دسته دوم به دسته‌بندی افراد متخصص، امتیازدهی به آن‌ها و شناسایی بهترین کاندیداها از بین موجود پرداخته می‌شود. در نهایت، دسته سوم شامل انجام تحلیل‌های دقیق‌تری بر روی داده‌های استخراج شده از دسته دوم با استفاده از متدهای علوم داده است. در این پژوهش، با تمرکز بر دسته سوم، داده‌های استخراج شده از کاربران بر اساس زمان دسته‌بندی شده و با استفاده از روش‌های آماری و احتمالاتی، پیش‌بینی‌هایی درباره تخصص‌هایی که یک متخصص ممکن است در آینده کسب کند ارائه می‌شود. این پیش‌بینی‌ها به سازمان‌ها و شرکت‌ها کمک می‌کند تا بهترین کاندیداها را بر اساس برنامه‌های آینده شغلی آن‌ها انتخاب کرده و با خطای کمتری در استخدام نیروی متخصص مواجه شوند.

واژگان کلیدی: بازیابی تخصص، بازیابی اطلاعات، تحلیل زمانی

فهرست مطالب

۱	مقدمه	۱
۳	ادبیات تحقیق	۲
۴	۱.۲ مدل های احتمالاتی	۴
۴	۲.۲ خوشه بندی	۴
۵	۳.۲ مدل های احتمالاتی بر پایه پیش بینی آینده	۵
۶	۴.۲ یادگیری ماشین	۶
۶	۵.۲ بازیابی سند	۶
۷	۶.۲ سری های زمانی	۷
۸	۷.۲ معیار های ارزیابی	۸
۱۰	۳ کار های مرتبط	۱۰
۱۸	۴ نتیجه گیری و کار های آینده	۱۸
۲۱	مراجع	۲۱
۲۲	واژه نامه انگلیسی به فارسی	۲۲
۲۳	واژه نامه فارسی به انگلیسی	۲۳

فهرست تصاویر

۱۶	یک نمونه از سری های زمانی	۱.۳
----	-------------------------------------	-----

فهرست جداول

فصل ۱

مقدمه

مسائل بازیابی تخصص به بررسی و تحلیل میزان تخصص افراد، حوزه های فعالیت آن ها و همچنین میزان عمیق بودن آن ها در هر حوزه تخصص می پردازد.

مطالعات پیشین در این حوزه به معرفی روش هایی میپردازد که با تحلیل متن هایی که هر کاربر در یک سری از وب سایت های مرتبط تولید کرده است . تخصص های آن کاربر را مشخص کرده و با تکنیک های خوشه بندی این مهارت ها را به دسته های مشخص تقسیم بندی میکند. از طرف دیگر مطالعاتی وجود دارد که میزان عمق دانش کاربران در هر حوزه تخصصی را مشخص میکند. و به این صورت مشخص میکند که کاربر از لحاظ میزان مهارت در حد یک کارآموز میباشد یا مهارت های عمیق تری دارد و میتواند آن را در سطوح بالاتر قرار داد.

در این مطلب تلاش شده در ادامه مطالعات پیشین روش هایی ارائه شود که از طریق آن با بررسی تغییرات تخصص افراد در حوزه های مختلف پیش بینی از تخصص هایی که فرد احتمالاً در آینده به سراغ آن ها خواهد رفت ارائه کند. این موضوع از چند جهت حائز اهمیت است. اول این که کسب و کار هایی که به دنبال جذب نیروی کارآموز و آموزش آن میباشند میتوانند با این تکنیک ها مسیر پیش بینی شده آینده آن فرد را پیدا کرده و از میان کاندید های موجود فردی را که مسیر یادگیری مهارت او تناسب بیشتری با نقشه راهی که آن کسب و کار برای آینده آن کارآموز در نظر گرفته مطابقت دارد جذب میکنند.

از جهت دیگر کسب و کار هایی که در نقشه راه آن ها بگونه ای است که در ماه ها و سال های آینده به سراغ یک تکنولوژی یا ابزار خاص خواهند رفت نیازی نیست که از ماه ها قبل فردی را با آن مهارت جذب کنند (این کار هزینه بالاتری برای کسب و کار دارد). آن ها با استفاده از این تکنیک فردی را که مهارت استفاده از ابزار هایی را دارد که کسب و کار در حال حاضر از آن استفاده میکند را دارد و پیشبینی مسیر یادگیری او به گونه ای است که در آینده در مهارتی که به کسب و کار اضافه میشود تخصص پیدا میکنند جذب میکنند.

علاوه بر دو مثال فوق میتوان مثال های دیگری نیز ارائه کرد که این راهکار فرآیند جذب و استخدام را بهینه میکند و هزینه نهایی آموزش نیروی جدید را برای کسب و کار ها کاهش میدهد. [۱]

فصل ۲

ادبیات تحقیق

۱.۲ مدل های احتمالاتی

مدل های احتمالاتی به منظور تحلیل و پیش بینی پدیده ها و داده ها با استفاده از اصول احتمال به کار می روند. این مدل ها از مفاهیم احتمالات برای مدل سازی عدم قطعیت و تغییرپذیری داده ها بهره می برند. از جمله کاربردهای مدل های احتمالاتی می توان به تشخیص الگوها، پیش بینی نتایج و تصمیم گیری تحت عدم قطعیت اشاره کرد. این مدل ها در زمینه های مختلف مانند یادگیری ماشین، پردازش زبان طبیعی و تحلیل داده کاربرد دارند. با استفاده از داده های تاریخی و توزیع های احتمالی، مدل های احتمالاتی می توانند نتایج آینده را با دقت بالاتری پیش بینی کنند.

۲.۲ خوشه بندی

خوشه بندی به عنوان یکی از تکنیک های مهم در حوزه بازیابی اطلاعات و به ویژه در بازیابی تخصص ها نقش بسزایی ایفا می کند. خوشه بندی فرآیندی است که طی آن، مجموعه ای از داده ها به گروه های مشابه و همگن تقسیم می شوند. این تکنیک با هدف دسته بندی داده ها به گونه ای انجام می شود که داده های موجود در هر خوشه بیشترین شباهت را به یکدیگر و کمترین شباهت را به داده های خوشه های دیگر داشته باشند. در حوزه بازیابی تخصص ها، خوشه بندی می تواند برای شناسایی و گروه بندی تخصص های مشابه استفاده شود. به طور مثال، می توان متخصصانی که در یک زمینه خاص فعالیت دارند را در یک خوشه قرار داد تا به این ترتیب، بازیابی اطلاعات مرتبط با تخصص ها به شکل کارآمدتر و دقیق تری انجام شود. این امر می تواند در سیستم های پیشنهاددهنده، جستجوهای تخصصی و تحلیل شبکه های اجتماعی به کار رود. روش های مختلفی برای انجام خوشه بندی وجود دارد که از جمله مهم ترین آن ها می توان به روش های مبتنی بر فاصله، و روش های سلسله مراتبی اشاره کرد. هر یک از این روش ها با توجه به ویژگی های داده ها و اهداف خوشه بندی، مزایا و محدودیت های خاص خود را دارند. در حوزه بازیابی اطلاعات، خوشه بندی علاوه بر دسته بندی تخصص ها، می تواند به بهبود دقت نتایج جستجو، کاهش پیچیدگی فضای جستجو و افزایش کارایی سیستم های بازیابی کمک کند. به عنوان مثال، با خوشه بندی مدارک و مقالات علمی، می توان به جستجوی موضوعات مرتبط پرداخت و نتایج دقیق تری را به کاربران ارائه داد.

در نهایت، می‌توان گفت که خوشه‌بندی به‌عنوان یکی از ابزارهای قدرتمند در تحلیل داده‌ها و بازیابی اطلاعات، نقش کلیدی در بهبود فرآیندهای بازیابی تخصص‌ها و مدیریت دانش ایفا می‌کند و می‌تواند به ارتقای سطح کیفی سیستم‌های اطلاعاتی منجر شود.

۳.۲ مدل‌های احتمالاتی بر پایه پیش‌بینی آینده

این مدل‌ها بر پایه تئوری احتمال و آمار بنا شده‌اند و با استفاده از داده‌های گذشته، سعی در پیش‌بینی وقایع و روندهای آینده دارند. در بازیابی تخصص‌ها، مدل‌های احتمالاتی می‌توانند برای پیش‌بینی نیازهای آینده به تخصص‌های خاص، شناسایی تغییرات در الگوهای تخصص و پیشنهاد تخصص‌های مورد نیاز در بازار کار استفاده شوند. یکی از کاربردهای مهم مدل‌های احتمالاتی در بازیابی تخصص‌ها، تحلیل پیش‌بینانه بر اساس داده‌های تاریخی است. این مدل‌ها با تحلیل داده‌های مربوط به فعالیت‌های گذشته متخصصان و روندهای شغلی، می‌توانند پیش‌بینی‌هایی در مورد آینده این تخصص‌ها ارائه دهند. به‌عنوان مثال، با استفاده از مدل‌های احتمالاتی می‌توان تخصص‌هایی که احتمال رشد بالایی دارند را شناسایی کرده و به کاربران پیشنهاد داد. از جمله روش‌های معروف در این حوزه، مدل‌های مارکوف، زنجیره‌های زمانی و تحلیل رگرسیون می‌باشند. مدل‌های مارکوف با استفاده از احتمال انتقال از یک حالت به حالت دیگر، روندهای آینده را پیش‌بینی می‌کنند. زنجیره‌های زمانی نیز با تحلیل داده‌های سری زمانی و بررسی الگوهای تکرار شونده، به پیش‌بینی آینده می‌پردازند. تحلیل رگرسیون نیز با استفاده از متغیرهای مستقل و وابسته، روابط بین داده‌ها را مدل‌سازی کرده و پیش‌بینی‌های دقیقی ارائه می‌دهد.

این مدل‌ها نه تنها در پیش‌بینی روندهای آینده، بلکه در بهبود سیستم‌های پیشنهاددهنده و جستجو نیز کاربرد دارند. به‌طور مثال، با پیش‌بینی نیازهای آینده به تخصص‌های خاص، سیستم‌های بازیابی می‌توانند به کاربران خود تخصص‌های مورد نیاز در آینده را پیشنهاد دهند و آن‌ها را برای بازار کار آینده آماده کنند. در مجموع، مدل‌های احتمالاتی با توانایی پیش‌بینی و تحلیل داده‌ها، نقش کلیدی در بهبود کارایی سیستم‌های بازیابی اطلاعات و تخصص‌ها ایفا می‌کنند و به بهینه‌سازی فرآیندهای تصمیم‌گیری و مدیریت دانش کمک می‌نمایند.

۴.۲ یادگیری ماشین

یادگیری ماشین به عنوان یکی از شاخه‌های پیشرفته و نوین علوم کامپیوتر، در سال‌های اخیر مورد توجه فراوانی قرار گرفته است. این فناوری با استفاده از الگوریتم‌ها و مدل‌های ریاضی قادر است به تحلیل و تفسیر داده‌ها پرداخته و الگوهای موجود در آنها را شناسایی کند. یکی از کاربردهای مهم یادگیری ماشین، پیش‌بینی آینده بر اساس داده‌های گذشته و حال است. در زمینه بازیابی اطلاعات، یادگیری ماشین می‌تواند نقش مؤثری در بهبود فرآیند بازیابی تخصص‌ها ایفا کند. بازیابی تخصص‌ها، که زیرمجموعه‌ای از بازیابی اطلاعات محسوب می‌شود، به معنای شناسایی و استخراج دانش و مهارت‌های افراد متخصص در یک حوزه خاص است. با استفاده از تکنیک‌های یادگیری ماشین، می‌توان الگوها و روابط پنهان میان داده‌های موجود را شناسایی و از آنها برای پیش‌بینی تخصص‌ها و مهارت‌های مورد نیاز در آینده استفاده کرد. مدل‌های یادگیری ماشین قادرند با تحلیل داده‌های مربوط به سوابق کاری، مقالات علمی، پروژه‌های تحقیقاتی و سایر منابع مرتبط، به شناسایی الگوهای تخصصی بپردازند. این مدل‌ها می‌توانند بر اساس داده‌های جمع‌آوری شده، نیازهای آتی بازار کار را پیش‌بینی کرده و به سازمان‌ها کمک کنند تا برای آینده آماده باشند. به عنوان مثال، با تحلیل داده‌های موجود، می‌توان پیش‌بینی کرد که کدام تخصص‌ها در سال‌های آتی بیشتر مورد نیاز خواهند بود و چه مهارت‌هایی باید تقویت شوند. به طور خلاصه، یادگیری ماشین با توانایی پیش‌بینی آینده و تحلیل داده‌های موجود، می‌تواند به عنوان ابزاری قدرتمند در بهبود فرآیند بازیابی تخصص‌ها و افزایش کارایی سازمان‌ها و نهادها عمل کند. این فناوری نه تنها به شناسایی تخصص‌های مورد نیاز کمک می‌کند، بلکه می‌تواند به توسعه و بهبود مهارت‌های فعلی نیز یاری رساند.

۵.۲ بازیابی سند

بازیابی سند به عنوان بخشی اساسی از علم اطلاعات، نقش مهمی در پژوهش‌های ادبیات موضوعی ایفا می‌کند. این فرآیند به معنای جستجو، انتخاب و بازیابی مستندات و اسناد مرتبط با یک موضوع خاص است. با استفاده از روش‌های پیشرفته بازیابی اطلاعات، می‌توان الگوها و روابط پنهان در بین متون ادبی را شناسایی کرده و به تحلیل دقیق‌تری از آنها پرداخت. بازیابی سند با استفاده از فنون یادگیری ماشین و پردازش زبان طبیعی،

می‌تواند به پیش‌بینی موضوعات و مسائل آینده در ادبیات موضوعی کمک کند. این روش‌ها با تجزیه و تحلیل متون موجود، می‌توانند الگوهایی را کشف کنند که در آینده به عنوان موضوعات جدید یا جنبه‌های نو ادبیات مطرح خواهند شد. به طور خلاصه، بازیابی سند با بهره‌گیری از رویکردهای پیشرفته علم اطلاعات و همچنین فنون یادگیری ماشین، می‌تواند در پژوهش‌های ادبیات موضوعی به بهبود فرآیند انتخاب و تحلیل متون کمک کند و به پیش‌بینی آینده این حوزه نیز کمک نماید.

۶.۲ سری‌های زمانی

سری‌های زمانی به عنوان یک روش تحلیلی در علوم داده‌ها، برای مدل‌سازی و پیش‌بینی رفتار آینده داده‌ها استفاده می‌شوند. این روش بر مبنای مشاهدات متوالی در طول زمان استوار است که می‌تواند شامل داده‌های مالی، آب و هوا، فروش و غیره باشد. هدف اصلی تحلیل سری‌های زمانی، شناسایی الگوها و روندهای موجود در داده‌ها و استفاده از آنها برای پیش‌بینی مقادیر آینده است. در تحلیل سری‌های زمانی، مفاهیم اساسی شامل روند، فصلی بودن و نویز هستند. روند به تغییرات بلندمدت در داده‌ها اشاره دارد، در حالی که فصلی بودن به الگوهای تکراری در بازه‌های زمانی مشخص اطلاق می‌شود. نویز نیز به نوسانات تصادفی و غیرقابل پیش‌بینی در داده‌ها گفته می‌شود. یکی از روش‌های متداول برای مدل‌سازی سری‌های زمانی، مدل‌های خودهمبستگی و میانگین متحرک است. در این مدل‌ها، مقادیر آینده به صورت تابعی از مقادیر گذشته و خطاهای پیش‌بینی قبلی تخمین زده می‌شوند. به عنوان مثال، مدل‌های خودهمبستگی مقادیر آینده را بر اساس همبستگی مقادیر گذشته پیش‌بینی می‌کنند، در حالی که مدل‌های میانگین متحرک از خطاهای پیش‌بینی‌های قبلی برای اصلاح مقادیر آینده استفاده می‌کنند. روش دیگر برای تحلیل سری‌های زمانی، تحلیل طیفی است که در آن سری‌های زمانی به اجزای فرکانسی تجزیه می‌شوند. این روش برای شناسایی دوره‌های مختلف نوسانات در داده‌ها مفید است. در حوزه بازیابی تخصص‌ها، تحلیل سری‌های زمانی می‌تواند به پیش‌بینی نیازهای آتی بازار کار و تخصص‌های مورد نیاز کمک کند. با تحلیل داده‌های تاریخی مربوط به استخدام‌ها، پروژه‌ها و مهارت‌های مورد تقاضا، می‌توان روندهای تغییرات در بازار کار را شناسایی کرده و برای برنامه‌ریزی‌های آتی استفاده کرد. به طور خلاصه، تحلیل

سری‌های زمانی با شناسایی الگوها و روندهای موجود در داده‌ها، ابزاری قدرتمند برای پیش‌بینی آینده و تصمیم‌گیری‌های مبتنی بر داده فراهم می‌کند. این روش به ویژه در حوزه‌هایی که نیاز به پیش‌بینی دقیق و برنامه‌ریزی بلندمدت دارند، کاربرد دارد.

۷.۲ معیارهای ارزیابی

ارزیابی مدل‌های بازایی اطلاعات در حوزه علوم کامپیوتر با استفاده از معیارهای مختلفی مانند دقت، بازآوری و میانگین دقت متوسط (MAP) انجام می‌شود که هر یک از این معیارها جنبه‌های متفاوتی از عملکرد مدل را اندازه‌گیری می‌کنند.

دقت (Precision) به معنای نسبت تعداد مدارک مرتبط بازایی شده به کل مدارک بازایی شده است. به بیان دیگر، دقت نشان می‌دهد که از میان مدارکی که توسط مدل بازایی شده‌اند، چند درصد از آن‌ها واقعاً مرتبط با پرسش کاربر بوده‌اند. به عنوان مثال، اگر مدل ۱۰ مدرک را بازایی کند و از این تعداد، ۷ مدرک مرتبط باشند، دقت مدل ۷۰ درصد خواهد بود. دقت بالا نشان می‌دهد که مدل قادر است نتایج دقیق‌تری را به کاربر ارائه دهد و خطای نوع اول (False Positive) کمتر است.

بازآوری (Recall) معیار دیگری است که تمرکز آن بر میزان بازایی مدارک مرتبط است. بازآوری به معنای نسبت تعداد مدارک مرتبط بازایی شده به کل مدارک مرتبط موجود در مجموعه است. به عبارت دیگر، این معیار نشان می‌دهد که مدل تا چه حد توانسته است تمامی مدارک مرتبط با پرسش کاربر را بازایی کند. برای مثال، اگر در مجموعه ۲۰ مدرک مرتبط وجود داشته باشد و مدل ۱۵ مورد از آن‌ها را بازایی کند، بازآوری ۷۵ درصد خواهد بود. بازآوری بالا نشان‌دهنده توانایی مدل در یافتن اکثر یا تمامی مدارک مرتبط است و نشان می‌دهد که خطای نوع دوم (False Negative) کمتر است.

میانگین دقت متوسط (MAP) یکی از جامع‌ترین معیارها برای ارزیابی مدل‌های بازایی اطلاعات است که هر دو جنبه دقت و بازآوری را در نظر می‌گیرد. MAP به این صورت محاسبه می‌شود که ابتدا برای هر پرسش دقت در هر رتبه‌ای که یک مدرک مرتبط بازایی می‌شود محاسبه می‌گردد. سپس دقت‌های میانگین برای هر پرسش محاسبه شده و در نهایت میانگین این دقت‌ها برای تمامی پرسش‌ها به عنوان MAP محاسبه می‌شود.

MAP می‌تواند به عنوان یک معیار کلی برای مقایسه عملکرد مدل‌های مختلف مورد استفاده قرار گیرد، زیرا هم دقت در رتبه‌های بالاتر و هم توانایی مدل در بازیابی مدارک مرتبط در طول تمامی رتبه‌ها را می‌سنجد. به دلیل این ویژگی‌ها، MAP اغلب در مقالات علمی و ارزیابی سیستم‌های بازیابی اطلاعات به کار می‌رود. با استفاده از این معیارها، محققان و مهندسان می‌توانند مدل‌های بازیابی اطلاعات را از جنبه‌های مختلف ارزیابی کرده و نقاط قوت و ضعف هر مدل را شناسایی کنند. این امر به بهبود سیستم‌های بازیابی اطلاعات و ارائه نتایج دقیق‌تر و کامل‌تر به کاربران کمک شایانی می‌کند. به طور کلی، درک صحیح از این معیارها و استفاده مناسب از آن‌ها می‌تواند تأثیر قابل توجهی در طراحی و بهینه‌سازی سیستم‌های بازیابی اطلاعات داشته باشد.

فصل ۳

کارهای مرتبط

به طور کلی کارهایی که در حوزه بازیابی تخصص انجام شده را میتوان به چهار دسته کلی تقسیم کرد این چهار دسته شامل موارد زیر میباشد.

- شناسایی منابع داده
- استخراج داده
- طبقه بندی تخصص ها
- خودکار سازی فرآیند استخراج و طبقه بندی

در این مطلب نتیجه کارهای پیشین در دسته سوم یعنی طبقه بندی تخصص ها مورد استفاده قرار گرفته و با استفاده از روش های خوشه بندی که در کارهای گذشته معرفی شده، حوزه های مهارتی آینده افراد شناسایی میشود. در این بخش به توضیح کارهای پیشین انجام شده در بخش سوم پرداخته میشود. [۲]

طبقه بندی تخصص ها

در این بخش از پایان نامه، مروری بر تحقیقات پیشین در زمینه بازنمایی تخصص انجام شده است. این حوزه با استفاده از تکنیک های مختلفی برای استخراج و نمایش تخصص افراد از داده های موجود، توسعه یافته است. بر اساس مرور انجام شده، این تحقیقات به سه جنبه اصلی تقسیم می شوند: روش های بازنمایی، پشتیبانی زمانی، و پشتیبانی معنایی. [۳][۴]

روش های بازنمایی تخصص

روش های مختلفی برای بازنمایی تخصص در متون علمی به کار گرفته شده اند که هر کدام بر اساس نوع داده ها و اهداف خاص تحقیقاتی طراحی شده اند. این روش ها را می توان به پنج دسته اصلی تقسیم کرد:

- **فراوانی اصطلاحات:** این روش بر مبنای تعداد دفعات تکرار اصطلاحات در متون علمی استوار است. در این رویکرد، هر واژه به عنوان یک بُعد در نظر گرفته می شود و بر اساس تعداد تکرار آن در متن، میزان تخصص فرد یا ارتباط متن با یک موضوع خاص ارزیابی می شود.

- **مدل‌های زبانی:** این مدل‌ها با استفاده از تحلیل احتمالی توزیع واژه‌ها در متون، یک مدل زبانی ایجاد می‌کنند که می‌تواند برای ارزیابی میزان شباهت متون جدید با متون قبلی به کار رود.
- **مدل‌های موضوعی:** این مدل‌ها توزیع احتمالی موضوعات در متون را تحلیل می‌کنند. یکی از معروف‌ترین روش‌ها در این دسته، Latent Dirichlet Allocation (LDA) است که برای شناسایی موضوعات اصلی در یک مجموعه داده به کار می‌رود.
- **گراف:** در این روش، داده‌های تخصص به صورت یک گراف نمایش داده می‌شوند. سپس با استفاده از الگوریتم‌های خاص مانند PageRank یا Random Walk، روابط و رتبه‌بندی تخصص‌ها تحلیل می‌شود.
- **روش‌های سفارشی:** این دسته شامل رویکردهای خاصی است که در آن‌ها از تکنیک‌های جایگزین برای بازنمایی تخصص استفاده می‌شود. به عنوان مثال، برخی مطالعات از داده‌های باز پیوندی (LOD) برای ساخت پروفایل‌های پژوهشگران استفاده می‌کنند.

پشتیبانی زمانی

تخصص افراد ممکن است با گذر زمان تغییر کند و این تغییرات می‌تواند به دلیل تغییر علاقه‌مندی‌ها یا تغییرات در حوزه‌های تحقیقاتی باشد. بنابراین، در تحلیل تخصص، باید زمان نیز به عنوان یک عامل مهم در نظر گرفته شود. تحقیقات در این زمینه به دو دسته اصلی تقسیم می‌شوند:

- **برش‌های زمانی:** در این رویکرد، تغییرات تخصص در طول زمان به صورت برش‌های زمانی مانند سال بررسی می‌شود. این روش به تحلیل تغییرات تدریجی در تخصص و پیش‌بینی علاقه‌مندی‌های آینده کمک می‌کند.
- **پیوسته:** این روش‌ها نیازی به برش‌های زمانی مشخص ندارند و تغییرات تخصص را به صورت پیوسته در طول زمان بررسی می‌کنند. این رویکردها با استفاده از برچسب‌های زمانی مرتبط با هر داده، تغییرات در تخصص افراد را تحلیل می‌کنند.

پشتیبانی معنایی

پشتیبانی معنایی به معنی استفاده از ساختارهای معنایی برای بهبود تحلیل و بازنمایی تخصص است. در حالی که بسیاری از مطالعات موجود از پشتیبانی معنایی استفاده نمی‌کنند، برخی از تحقیقات از هستی‌شناسی‌ها (Ontology) و پایگاه‌های دانش (Knowledge Base) برای این منظور بهره می‌برند. این ابزارها به تحلیل بهتر مفاهیم موجود در متون و ارتباطات معنایی بین آن‌ها کمک می‌کنند.

خوشه بندی تخصص ها

در این بخش کارهای پیشین با تمرکز روی تحلیل متن پاسخ کاربران در وبسایت‌ها سعی در استخراج تخصص آن‌ها و سپس قرار دادن هر پاسخ در یک حوزه مهارتی می‌باشد. همچنین با این روش حوزه‌های مهارتی افراد شناسایی می‌شود.

درخت موضوعی

در این بخش به توضیح سلسه مراتب موضوعاتی که در این مطلب آمده است و در زمینه آن کارهایی انجام شده است پرداخته می‌شود.

بازیابی اطلاعات

والد تمام برگ‌های درخت موضوعی بازیابی اطلاعات می‌باشد. بازیابی اطلاعات یکی از حوزه‌های مهم در علوم کامپیوتر است که به مطالعه روش‌ها و تکنیک‌های جستجو و بازیابی اطلاعات از مجموعه‌های عظیمی از داده‌ها می‌پردازد. با توجه به رشد روزافزون تولید داده، سیستم‌های بازیابی اطلاعات نقش مهمی در ارائه نتایج مرتبط و دقیق به کاربران ایفا می‌کنند. این سیستم‌ها در موتورهای جستجوی وب، پایگاه‌های داده و بسیاری از کاربردهای دیگر مورد استفاده قرار می‌گیرند. مباحث مهم در این حوزه شامل مدل‌های بازیابی، الگوریتم‌های ایندکس‌سازی، پردازش زبان طبیعی و یادگیری ماشین است. چالش‌های اصلی در بازیابی اطلاعات عبارتند از: حجم عظیم داده‌ها، تنوع داده‌ها، ابهام و چند معنایی پرس‌وجوها و تغییرات مداوم در داده‌ها. با مطالعه و بررسی

این حوزه، می‌توان راهکارهای نوینی برای بهبود کارایی سیستم‌های جستجو و کشف اطلاعات ارائه داد.

بازیابی تخصص

بازیابی تخصص فرزند مستقیم بازیابی اطلاعات میباشد. بازیابی تخصص شاخه‌ای از علم اطلاعات است که به یافتن افراد یا منابعی که دارای دانش و مهارت‌های خاص در یک حوزه مشخص هستند، می‌پردازد. با توجه به پیچیدگی روزافزون مسائل و نیاز به متخصصین در حوزه‌های مختلف، بازیابی تخصص به یکی از چالش‌های مهم در سازمان‌ها و جوامع تبدیل شده است.

سیستم‌های بازیابی تخصص با استفاده از تکنیک‌های مختلف مانند پردازش زبان طبیعی، یادگیری ماشین و شبکه‌های اجتماعی، تلاش می‌کنند تا افراد با تخصص مناسب را شناسایی و به یکدیگر متصل کنند. این سیستم‌ها با تحلیل رزومه‌ها، مقالات علمی، پروژه‌ها و سایر اطلاعات مرتبط، پروفایل‌های تخصصی افراد را ایجاد کرده و سپس با توجه به نیازهای یک پروژه یا سوال خاص، افراد مناسب را پیشنهاد می‌دهند.

مباحث مهم در حوزه بازیابی تخصص شامل مدل‌های نمایش دانش، الگوریتم‌های تطبیق، شبکه‌های اجتماعی علمی و ارزیابی سیستم‌های توصیه‌گر است. چالش‌های اصلی در این حوزه عبارتند از: تعریف دقیق تخصص، ابهام در توصیف نیازها، پویایی دانش و مهارت‌ها و حفظ حریم خصوصی اطلاعات شخصی.

طبقه بندی تخصص

این بخش فرزند مستقیم بازیابی تخصص میباشد که در مقدمه این فصل به آن اشاره شده است.

خوشه بندی

خوشه‌بندی به منظور گروه‌بندی داده‌های مشابه با یکدیگر به کار می‌رود. در واقع، خوشه‌بندی به دنبال یافتن الگوهای پنهان در داده‌هاست و به این ترتیب، به ما کمک می‌کند تا داده‌های بزرگ و پیچیده را بهتر درک کنیم و از آن‌ها برای تصمیم‌گیری بهتر استفاده کنیم. برای مثال، می‌توانیم از خوشه‌بندی برای دسته‌بندی مشتریان یک فروشگاه بر اساس رفتار خرید آن‌ها استفاده کنیم یا مقالات علمی را بر اساس موضوع آن‌ها گروه‌بندی کنیم. الگوریتم‌های مختلفی برای خوشه‌بندی وجود دارد که هر کدام برای نوع خاصی از داده‌ها مناسب هستند.

از جمله این الگوریتم‌ها می‌توان به k -means، سلسله مراتبی، مبتنی بر چگالی و مبتنی بر مدل اشاره کرد. انتخاب الگوریتم مناسب به عوامل مختلفی مانند نوع داده‌ها، تعداد خوشه‌ها و هدف از خوشه‌بندی بستگی دارد. خوشه‌بندی کاربردهای بسیار گسترده‌ای در حوزه‌های مختلف مانند تجارت، پزشکی، بیوانفورماتیک و علوم اجتماعی دارد. در این درخت موضوعی خوشه‌بندی یکی از فرزندان مستقیم بحث طبقه‌بندی تخصص می‌باشد.

مدل‌های احتمالاتی

مدل‌های احتمالاتی نیز فرزند گره بازیابی تخصص می‌باشند البته این مدل‌های در حوزه‌های دیگر نیز کاربرد گسترده‌ای دارند می‌باشند ولی در اینجا کاربرد آن در مباحث مربوط به بازیابی اطلاعات و مخصوصاً بازیابی تخصص مد نظر می‌باشد.

تحلیل زمانی

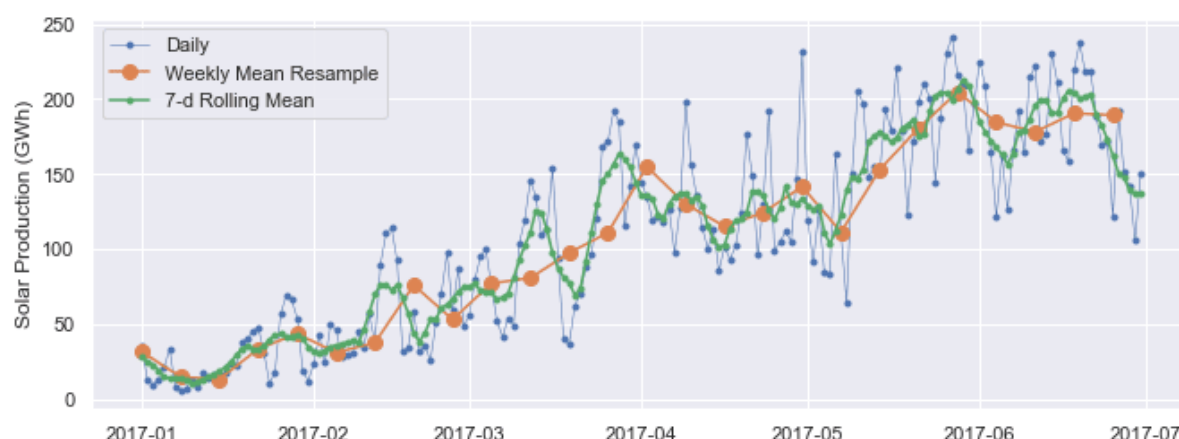
در این حوزه تغییرات داده در طول زمان بررسی می‌شود. این بخش فرزند بخش طبقه‌بندی تخصص می‌باشد. به گونه‌ای که تغییرات تخصص افراد در طول زمان بررسی می‌شود

سری‌های زمانی

این سری‌ها داده‌های ساختاریافته‌ای هستند که یکی از مشخصه‌های آن‌ها زمان می‌باشد. در نتیجه می‌توان این داده‌ها را در محور زمان مرتب‌سازی و تحلیل کرد.

پیشبینی آینده

در این حوزه سعی می‌شود با تحلیل داده‌های گذشته (معمولاً داده‌های سری‌های زمانی) تخمین نسبتاً دقیقی از آینده ارائه کرد. روش‌های انجام این پیشبینی معمولاً به دو دسته کلی تقسیم می‌شوند که شامل روش‌های مبتنی بر یادگیری ماشین و روش‌های آماری می‌باشد. هر دو این روش‌ها در حوزه بازیابی تخصص می‌توانند مفید واقع شوند ولی به صورت کلی پیاده‌سازی روش‌های آماری به مراتب ساده‌تر از روش‌های مبتنی بر یادگیری ماشین می‌باشد.



شکل ۱.۳: یک نمونه از سری های زمانی

روش های ارزیابی

قبل از شروع یافتن روش مناسب برای تخمین تخصص آینده افراد لازم است روشی برای ارزیابی این تخمین وجود داشته باشد. ساده ترین روش این است که داده های مورد آزمون به دو دسته تقسیم شده، دسته قدیمی تر به عنوان داده هایی برای تخمین و دسته دوم داده هایی هستند که میزان نزدیک بودن تخمین تولید شده از داده های بخش اول میتواند دقت روش را ارزیابی کند. البته این روش همیشه کارساز نخواهد بود، زیرا در برخی مواقع پس از تخمین داده های جدیدی وجود ندارد که بتوان دقت روش را ارزیابی کرد، در این صورت لازم است از روش های پیچیده تری استفاده شود.

از ساده ترین روش های ارزیابی میتوان سه مورد زیر را نام برد.

یادآوری (Recall)

یادآوری یکی از معیارهای ارزیابی مدل های یادگیری ماشین است که به درصد نمونه های مثبت درست پیش بینی شده نسبت به کل نمونه های مثبت اشاره دارد. به عبارت دیگر، یادآوری نشان می دهد که مدل تا چه حد قادر است تمام نمونه های مثبت موجود در داده ها را شناسایی کند. این معیار در کاربردهایی که تشخیص همه موارد مثبت اهمیت دارد، بسیار مهم است؛ مثلاً در شناسایی بیماری ها که مهم است هیچ موردی نادیده گرفته نشود. یادآوری بالا نشان دهنده این است که مدل موفق به شناسایی تعداد زیادی از نمونه های مثبت شده، اما ممکن است برخی از نمونه های منفی را به اشتباه مثبت پیش بینی کرده باشد. در تعادل میان یادآوری و دقت،

افزایش یادآوری معمولاً به کاهش دقت منجر می‌شود. این معیار به ویژه در مسائلی که خطاهای نوع دوم (عدم شناسایی موارد مثبت) حیاتی است، اهمیت دارد.

نقشه (MAP)

میانگین دقت ترتیبی یک معیار ارزیابی مدل‌های بازبازی اطلاعات و یادگیری ماشین است که کیفیت رتبه‌بندی نتایج را می‌سنجد. این معیار برای مدل‌هایی که چندین خروجی به صورت مرتب شده تولید می‌کنند، اهمیت دارد. نقشه به صورت میانگین دقت در سطوح مختلف رتبه‌بندی محاسبه می‌شود و نشان می‌دهد که نتایج بالاتر رتبه به چه میزان مرتبط هستند. به عبارت دیگر، نقشه به ارزیابی توانایی مدل در مرتب‌سازی صحیح نتایج مرتبط کمک می‌کند. این معیار بیشتر در سیستم‌های پیشنهاددهنده، موتورهای جستجو و کاربردهایی که رتبه‌بندی نتایج اهمیت دارد، استفاده می‌شود. نقشه بالاتر نشان‌دهنده کیفیت بهتر رتبه‌بندی مدل است. این معیار با محاسبه میانگین دقت در هر نقطه که یک نمونه مثبت رتبه‌بندی شده، به دست می‌آید. در نهایت، نقشه ابزاری است برای سنجش کلی عملکرد یک مدل در زمینه رتبه‌بندی.

دقت (Precision)

دقت یکی از معیارهای اساسی ارزیابی مدل‌های یادگیری ماشین است که به نسبت تعداد نمونه‌های مثبت درست پیش‌بینی شده به کل نمونه‌هایی که به عنوان مثبت پیش‌بینی شده‌اند اشاره دارد. به عبارت دیگر، دقت نشان می‌دهد که مدل تا چه حد در پیش‌بینی‌های مثبت خود درست عمل کرده است. دقت بالا به معنای این است که تعداد پیش‌بینی‌های نادرست (مثبت کاذب) کم است و مدل بیشتر بر پیش‌بینی‌های صحیح متمرکز است. این معیار در مواقعی که هزینه اشتباهات مثبت کاذب بالاست، اهمیت ویژه‌ای دارد؛ مثلاً در مواردی که تشخیص اشتباه یک تهدید می‌تواند منجر به اقدامات غیرضروری شود. در مقابل، افزایش دقت معمولاً به کاهش یادآوری منجر می‌شود، زیرا مدل تنها مواردی را به عنوان مثبت پیش‌بینی می‌کند که از صحت آن‌ها مطمئن است. دقت به عنوان یکی از معیارهای مهم در تعیین کارایی مدل‌های یادگیری ماشین شناخته می‌شود و در تعادل با یادآوری به کار می‌رود تا بهترین عملکرد مدل مشخص شود.

فصل ۴

نتیجه‌گیری و کارهای 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 بررسی 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 در نظر 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 کارها 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 دقیق تر میباشد.

نتیجه گیری

به طور کلی مسئله بازیابی نیروی متخصص یک مسئله نسبتاً 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 آن ها برای یافتن بهترین کاندید یک موقعیت شغلی میباشد. ضعف اصلی روش های موجود این است که این است که این روش ها معمولاً پارامترهای کمی را برای بررسی 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] K. Balog and Y. Fang, "Expertise retrieval (foundations and trends in information retrieval)," *now*, 2021.
- [2] rodrigo gonçaves and carina friedrich dorneles, "Automated expertise retrieval : A taxonomy-based survey and open issues," *ACM Computing Surveys (CSUR)*, vol.52, no.2-3, pp.1-30, 2019.
- [3] P. Rostami and M. Neshati, "Automated expertise retrieval : A taxonomy-based survey and open issues," *Expert Systems with Applications*, 2021.
- [4] P. Rostami and M. Neshati, "T-shaped grouping : Expert finding models to agile software teams retrieval," *Expert Systems with Applications*, 2019.

واژه‌نامه انگلیسی به فارسی

map	میانگین دقت ترکیبی
Precision	دقت
recall	یادآوری

واژه‌نامه فارسی به انگلیسی

clustering	خوشه بندی
information retrival	بازیابی اطلاعات
probabilistic model	مدل های احتمالاتی
mashin learning	یادگیری ماشین
experties	متخصص
feature forecast	پیشبینی آینده
time series	سری های زمانی
data extraction	استخراج داده
profile	نمایه

Abstract:

Expertise retrieval is a research topic within the field of information retrieval that aims to identify specialists and analyze large-scale data, which often appears in the form of unstructured texts. This process helps organizations and companies easily find the specialized workforce they need. Generally, expertise retrieval can be divided into several categories. The first category involves data extraction and preparation, where data is typically extracted from online forums for research purposes. The second category includes the classification of specialists, scoring them, and identifying the best candidates from the available pool. Finally, the third category involves performing more detailed analyses on the data extracted in the second category using data science methods. In this study, with a focus on the third category, the extracted user data is categorized based on time, and by applying a series of statistical and probabilistic methods, predictions are made about the potential expertise a specialist may acquire in the future. These predictions assist organizations and companies in selecting the best candidates based on their future career plans and minimizing errors in hiring specialized personnel.

Keywords: Information retrieval, expertise retrieval, time series



Shahid Beheshti University

Faculty of Computer Science & Engineering

feature forcast methods in experties retrival

By

sina mohammadi

A THESIS SUBMITTED
FOR THE DEGREE OF
MASTER OF SCIENCE

Supervisor :

Dr. Mahmood Neshati

2024