ترجمه فا
TarjomeFa.Com

ارائه شده توسط:

سایت ترجمه فا

مرجع جدیدترین مقالات ترجمه شده

از نشریات معتبر

سیستم GeoSRS: یک سیستم پیشنهاد دهنده اجتماعی هیبرید برای داده های مبتنی بر موقعیت جغرافیایی کاربر

چکیده:

ما سیستم GeoSRS را ارائه داده ایم که یک سیستم پیشنهاد دهنده هیبرید برای یک شبکه اجتماعی مشهور مبتنی بر موقعیت کاربر یا LBSN می باشد، که در آن کاربران قادرند مطالب مروری کوتاهی درباره محلهای مورد علاقه ای که دیدن کرده اند بنویسند. با استفاده از تکنیک های داده کاوی متن پیشرفته، سیستم ما موقعیت هایی را به کاربران توصیه کرده اند تا به عنوان مجموعه کاملی از مرور های متنی علاوه بر محل جغرافیایی شان استفاده بشود. برای ارزیابی سیستم خود، ما مجموعه داده های خودمان را با خزش در شبکه اجتماعی Foursquare جمع آوری کرده ایم. برای انجام موثر چنین کاری، ما استفاده از نسخه موازی تکنیک Quadtree را پیشنهاد کرده ایم که می تواند برای خزش/کاوش منابع توزیع شده فضایی دیگر کاربرد داشته باشد. سرانجام اینکه، ما عملکرد GeoSRS را روی مجموعه داده های جمع آوری شده مان مطالعه کرده ایم و نتیجه گرفته ایم که با ترکیب تحلیل گرایشات و احساسات و مدلسازی متنی، GeoSRS پیشنهادات صحیح تری را بوجود می آورد. عملکرد این سیستم جوری بهبود یافته است که مرورهای بیشتری در دسترس خواهند بود، و استفاده از تکنیک های خزش در مقیاس بزرگ را نظیر Quadtree بیشتر ترغیب خواهد کرد.

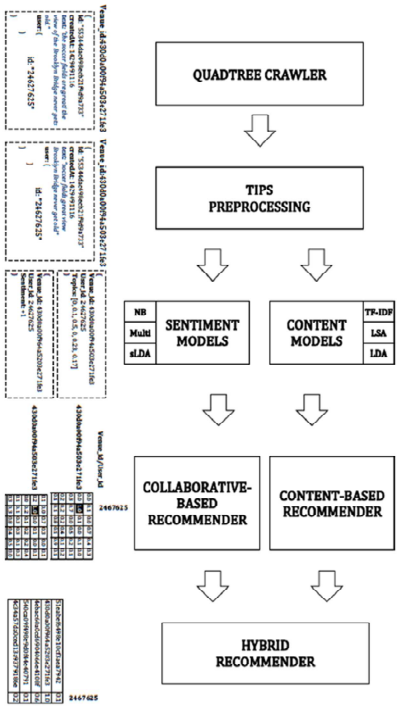
کلیدواژه ها: سیستم های پیشنهاد دهنده، متن کاوی، تکنیک Quadtree، خزش، شبکه های اجتماعی، شبکه اجتماعی مبتنی بر موقعیت کاربر.

۱-مقدمه

تکثیر تکنولوژی های ارتباطات موبایل و GPS باعث شده کاربران کلان داده های شناسایی جغرافیایی را به رسانه های اجتماعی مختلفی اضافه نمایند از جمله عکس، مرور متنی، یا ویدئو که چند نمونه از بسیار نمونه های دیگر می باشند. شبکه های اجتماعی مبتنی بر موقعیت کاربر یا LBSN روابط کاربر (بخش اجتماعی) و اطلاعات فضایی جغرافیایی (بخش مبتنی بر موقعیت کاربر) را به شکل یک شبکه منفرد با هم ترکیب کرده اند. با درنظرگیری

موقعیت فیزیکی کاربران، LBSNها فاصله دنیای فیزیکی و جوامع مجازی را نظیر Foursquare و Facebook یا Twitter را پر می کنند.

استفاده گسترده از این محلهای شبکه سازی اجتماعی باعث شده که اینها منابع ارزشمند اطلاعاتی بشوند. ولیکن، حجم زیاد داده هایی که از این محلها جریان می یابد، کار را حتی از لحاظ یک کاربر منفرد، به طور روزافزونی مشکل می سازد که انسانها همه این اطلاعات را راهگیری کنند. ازا ینرو بیشتر سایت های شبکه اجتماعی نوعی از سیستم توصیه اجتماعی SRS را اجرا می نمایند: برای مثال، توییتر می گوید که چه کسی را دنبال کنید، فیس بوک پست ها را در قسمت دیوار کاربران فیلترسازی و اولویت بندی می کند.



شکل ۱-سیستم اطلاعات مبتنی بر موقعیت کاربر کاملا یکپارچه سازی شده. این سیستم شامل مودولهایی برای خزش، پیش پردازش داده های حاصل از خزش، مدلسازی و پیشنهادات می باشد. بخش پایینی این تصویر نشاندهنده نمونه هایی از حجم داده هایی است که هر مودول ایجاد کرده است.

و Foursquare رفتن به مکانهایی را پیشنهاد می کند. وقتی محتوای اجتماعی از لحاظ جغرافیایی نشاندار بشود، به شدت لازم می شود که کاربر و مکان یابی گزینه در پارادیگم پیشنهادات درنظر گرفته شود.

پیشنهادات مبتنی بر موقعیت کاربر یک کاربر بینظیری در LBSNها دارد و اساسا از سیستم های پیشنهاددهنده معمولی متفاوت است از این لحاظ که سیستم های معمولی خصوصیات فضایی کاربران و گزینه ها را درنظر نمی گیرند. بعلاوه، پیشنهادات مبتنی بر موقعیت کاربر در بالای LBSNها ممکن است از تعامل میان سه لایه تشکیل دهنده یک LBSN یعنی کاربر، موقعیت جغرافیایی و لایه محتوایی بهره مند گردد.

در این مقاله، ما یک سیستم کاملا یکپارچه ای را برای بازیابی اطلاعات داده های مبتنی بر موقعیت جغرافیایی کاربر و سیستم پیشنهاد دهنده مبتنی بر موقعیت انتها به انتها پیشنهاد داده ایم که برای شبکه اجتماعی مشهور Foursquare مناسب می باشد. ولیکن خواننده باید توجه داشته باشد که روش کار ما برای شبکه های اجتماعی دیگر هم کاربرد دارد که حاوی مرورهای مرجع زمانی مبتنی بر موقعیت جغرافیایی میباشد و با اینحساب در مقاله مان، از این حقیقت جدا می شویم که داریم از یک سایت خاصی استفاده می کنیم.

ما براین باوریم که سیستم های پیشنهاد دهنده کاملا عملیاتی در بالای LBSNها نیاز به طراحی های انتهاب به انتها دارند، که قادر به اجرای بازیابی داده ها از شبکه های اجتماعی و پاکسازی سروصدا و داده های تکراری، و استخراج ویژگی های مرتبط و نه دست کم اجرای پیشنهادات می باشند.

سیستم اطلاعات مبتنی بر موقعیت کاربر که ما پیشنهاد داده ایم، در شکل ۱ خلاصه سازی شده است. این سیستم مرورهای کوتاه را همراه با محلهای مکان یابی شده جغرافیایی شان و شناسایی مرورکننده ها به عنوان اساسی برای پیشنهادات بازیابی می کند. در Foursquare، کاربران قادر به بررسی محلهای مورد علاقه (مسیرها)، نوشتن مرورهای کوتاه (نکات مهم) برای محلهایی که آنها را بررسی کرده اند، و به اشتراک گذاری اطلاعات با کاربران درون شبکه اجتماعی شان می باشند. برای کار خزش سازی محل ها، کاربران و نکات مهم، ما مجبور به استفاده از APIی Foursquare گردیدیم که یک سطح مشترک با محدودیت های مقدار اطلاعاتی است که فرد می پرسد و مقدار درخواست هایی است که فرد دارد. از اینرو، لازم بود یک مکانیسم خزش سازی ابداع گردد که استفاده بهینه از سوالات در دسترس ما قرار گیرد. برای دستیابی به این کار، ما یک نسخه موازی الگوریتم Quadtree را طراحی کرده ایم، که خیلی مناسب خزش سازی محلهایی می باشد که به لحاظ فضایی توزیع یافته اند، درحالیکه در عین حال ظرفیت پذیرش قابل ملاحظه ای را کسب کرده اند. ما دریافته ایم که خزش سازی همه محلها از نواحی شهری بزرگ مانند Mexico D.F. یا نیویورک در زمان منطقی با الگوریتم مطرح شده Quadtree

امکانپذیر بوده است. ما بکارگیری الگوریتم Quadtree موازی را برای این مسئله در مقاله خود به عنوان یک کاری ارزشمند می دانیم و براین باوریم که مسائلی که نیازمند وسایل فضایی کاوشی (سنسورها) می باشند می توانند همچنین از این امر بهره مند شوند.

برای انجام پیشنهادات سیستم ما استفاده زیادی از مرورهای کاربری (نکات مهم) نموده است. برای استخراج اطلاعات معنی دار از این مرورهای به شکل ازاد، سیستم GeoSRS متکی بر بسیاری تکنیک های پیشرفته برای متن کاوی و تحلیل احساسات و گرایشات می باشد، که از لحاظ صحت توصیه ارزیابی می شوند و متکی بر تکنیک هایی است که اجرای عالی دارند انتخاب می شوند که در GeoSRS استفاده بشوند. نقش مرتبط دیگری از مقاله ما همان افزایش صحت هنگام ترکیب گرایشات مرور و محتوا به شکل تنظیم پیشنهاددهنده توزین شده ساده ولی تا اندازه ای موثر می باشد. گرایشات اشاره به عقاید جهانی دارد که در مرور مربوطه بازتاب یافته است (مثبت، منفی یا خنثی) درحالیکه محتوا نشان دهنده موضوعاتی است که مرور مطرح می سازد. این امر باعث تقویت این ایده می شود که انتخابهای مبتنی بر مرور خالص و محض صرفا براساس عقایدی نیست که در یک مرور کوتاه بازتاب یافته است (خدمات خیلی کند می باشند)، بلکه براساس محتوای مربوط به کاربر است (این یک نوع محل کار به جای کافی شاپ است).

برای ارزیابی سیستم خود، ما مجموعه داده های خودمان را از رستورانها و نکاتی از ناحیه منهتن در نیویورک جمع آوری کرده ایم. ما منهتن را به دلیل تراکم بالای محله ها و تعداد کاربران فعال برای روایی سازی هم قابلیت معیاربندی خزشگر Quadtree و هم کارایی و پوشش سیستم پیشنهاد دهنده انتخاب کرده ایم. سیستم پیشنهاددهنده به شکل سنجشهای صحت بازیابی (عملکرد) در عوض سنجشهای صحت آماری ارزیابی می شود چون ما قصد داشتیم درجه بندی فردی محل ها را پیشگویی کنیم و ترتیب بندی نسبی در میان آنها را انجام دهیم، یک روش ارزیابی را مطرح کردیم که در آن ما به لحاظ تاریخی مجموعه داده های نکات مهم را به شکل آموزشی و تستی تقسیم بندی کرده ایم. نکات مهم تستی به شکل حقیقت زمینه ای برای ارزیابی مقایسه ای پیشنهادات درنظر گرفته شده اند. دست آخر اینکه، تنظیم پیشنهاد دهنده هیبرید توزین شده ساده که در GeoSRS بکار گرفته شده است طبق پیکربندی های پیشرفته دیگر مانند مدلهای کلان سطح و ابشاری مقایسه شده اند.

برای خلاصه بندی، این مقاله یک سیستم پیشنهاد دهنده هیبرید را برای یک شبکه اجتماعی مبتنی بر محل مطرح کرده است که به طور بینظیری طبق مرور متن ساخته شده است. نقش اصلی ما عبارت است از:

۱-استفاده از یک نسخه موازی تکنیک Quadtree به شکل راهکار پایه برای خزش سازی داده های توزیع شده فضایی

۲-استفاده از تحلیل احساسات و گرایشات روی مرورهای متنی برای تضمین منبع برای فیلترسازی جمعی

۳-استفاده از مرورهای جمع بندی شده توسط کاربر و محل ایجاد اطلاعات نمایه سازی شده برای پیشنهادات مبتنی بر محتوا

۴-استفاده از یک تکنیک هیبریدسازی ساده ولیکن قوی برای بهبود عملکرد توصیه.

۵-جمع بندی همه اینها به شکل یک سیستم اطلاعات کارامد

۶-ارزیابی و مقایسه GeoSRS علیه سایر سیستم های هیبرید پیشرفته از لحاظ ارقام IR

بقیه مقاله به ترتیب ذیل سازماندهی شده است: بخش ۲ نمایانگر یک مرور اجتماعی کار مربوطه در توصیه اجتماعی در بالای LBSNها ، سیستم های تحلیل گرایشات و محتوایی، و تکنیک های پیشنهادی هیبرید عمومی می باشد. بخش ۳ مطرح کننده یک تکنیک کارامد و موازی برای بازیابی منابع داده توزیع شده فضایی می باشد. سیستم GeoSRS ما در بخش ۴ توضیح داده شده است. بخش ۵ مجموعه داده های مروری با تعیین موقعیت جغرافیایی را از Foursquare با سیستم کارامد GeoSRS در کنار هم آورده است تا تنظیمات احتمالی مختلف را با استفاده از یک روش اجرای ارزیابی آفلاین ارزیابی نماید. دست اخر اینکه بخش ۶ نتیجه گیری هایی را از کار ما گرفته و شامل رهنمودهایی برای کار اتی می باشد.

۲-کارهای مربوطه

سیستم پیشنهاد دهنده مخالف این مقاله جزو رده سیستم های پیشنهاد دهنده اجتماعی مبتنی بر موقعیت یابی کاربر با استفاده از تحلیل گرایشات و محتوایی متن همراه با تکنیک های فیلترسازی جمعی قرار می گیرد که منجر به یک سیستم توصیه دهنده هیبرید می شود. برای اینکه سیستم خود را در زمینه متون معین شده قرار دهیم به اختصار هر یک از این حیطه های تحقیقاتی را که به کارمان مربوط می شود بررسی می کنیم.

سیستم های پیشنهاد دهنده اجتماعی (SRS): SRS به شکل کاربرد سیستم های پیشنهاددهنده به شبکه های اجتماعی ناشی می شود، هر چند آنها اخیرا به شکل رشته جداگانه ای به تنهایی پذیرفته شده اند. مع ذلک، محققان تنظیمات پیشنهاد دهنده تازه ای را براساس محتوای اجتماعی برای پنج الی ده سال اخیر مطرح کرده اند.

پارادیگم های پیشنهاددهنده کلاسیک مانند پیشنهاددهنده های مبتنی بر محتوا با افزودن اطلاعات دوست یابی به تابع عینی فاکتورگیری ماتریس با پیکربندی به اصطلاح پیشنهاددهنده مبتنی بر محتوای اجتماعی (SoCo) تقویت سازی شده اند. فیلترسازی مبتنی بر همکاری نیز با اضافه کردن اطلاعات اجتماعی برای بهبود شناسایی مجاورت کاربر و سروکاریابی با نایابی داده ها بازتعریفی می شوند. طبق رفرانس ۱۵، فیلترسازی اجتماعی از فیلترسازی مبتنی بر همکاری بهتر عمل می کند. سیستم پیشنهاد دهنده مبتنی بر اعتماد ایده فیلترسازی اجتماعی را حتی بیشتر تصفیه کرده و چندین ایده تکثیر اعتماد یا شهرت را از طریق شبکه کاربران تعریف کرده است. مثالهایی از این دست عبارتند از FeedbackTrust و TrustWalker هر چند بسیاری مثالهای دیگر هم درکار است. FeedbackTrust پیشنهادات مبتنی بر کاربر را با تقویت سازی شباهت های کاربر با شباهت های مبتنی بر اعتماد بهبود می دهد. TrustWalker پیشنهادات مبتنی بر گزینه را با ترکیب نتیجه راهپیمایی تصادفی تکراری در شبکه کاربران به هم متصل انجام می دهد.

سایرین مکانیسم هایی را برای لحاظ کردن تاثیرات بین فردی در سیستم های پیشنهاد دهنده معمولی مطرح کرده اند با این بحث که تاثیرات بین فردی یک نقش حیاتی را در این سناریو ایفا می کند. اصطلاح تعیین مقررات اجتماعی برای اشاره به استفاده از یک تعیین مقررات براساس محتوای اجتماعی ساخته شده است. در زمینه پیشنهادات گروهی، که شامل دادن پیشنهاداتی به یک گروه از افراد براساس علایق آنها می شود، کار Gartrell و همکارانش مزیت های لحاظ ساختار اجتماعی را در سیستم پیشنهاددهنده روشن کرده اند.

چندین SRS عملی را می توان در متون علمی یافت که برای مجزاسازی موقعیت های پیشنهاد دهنده استفاده می شوند. برای مثال، Diaby و همکارانش یک سیستم پیشنهاد دهنده شغلی مبتنی بر شبکه اجتماعی انلاین را برای استخدام کنندگان توضیح می دهد. Tu و همکارانش یک سیستم پیشنهاددهنده انلاین را برای قرارهای دوستانه ارائه داده اند که ترجیحات کاربر را از روی مدل تخصیص هدایت پذیری نهفته یا LDA تشخیص می دهد. Xia و

همکارانش یک سیستمی را مطرح کرده اند که پیوندهای اجتماعی را در حیطه مقالات علمی برای توصیه مقالات علمی به کاربران به حساب اورده است.

سیستم های پیشنهاددهنده اجتماعی مبتنی بر موقعیت کاربر: سیستم پیشنهاددهنده می تواند باز با بهره گیری از داده های موقعیت جغرافیایی برای لحاظ بعد فضایی بهبود یابد. چنین سیستم هایی کلا به سیستم های پیشنهاددهنده اجتماعی مبتنی بر موقعیت جغرافیایی اشاره دارند. مثالهایی در این راستا از تحقیقات همان Berjani & Strufe و Yang و همکارانش ، Ye و همکارانش و Zheng و همکارانش میب اشد. برای مقدمه ای بر سیستم های پیشنهاددهنده LBSN و تازه های پیشرفتها به اموزش رفرانس ۵۴ و کتاب درسی تازه منتشره نوشته Symeonidis در رفرانس ۴۸ رجوع شود. در مقایسه با مقاله Yang و همکارانش، ما از منبع داده های یکسانی استفاده کرده ایم و به دلیل شباهت نقش ها به موجب آن نقش های خود را با مال آنها با هم ارائه کرده ایم. مقاله Yang و همکارانش یک سیستم پیشنهادات مبتنی بر موقعیت جغرافیایی را مطرح نموده است که از نکات مهم موقعیت جغرافیایی از Foursquare هراه با اطلاعات بررسی برای بهبود عملکرد پیشنهادات استفاده برده است. سیستم پیشنهاددهنده هیبرید ما یعنی GeoSRS از این کار در جنبه های مختلف متفاوت است. درحالیکه مقاله Yang و همکارانش تنها از اطلاعات گرایشات حاصل از نکات مهم بهره برده است، همچنین ما ساختار موضوعات را از روی آنها برای مدلسازی مشخصات کاربر و محل استخراج کرده ایم. بعلاوه، نویسندگان براساس طرح پیشنهادات روی یک فاکتورگیری احتمال گرایانه ماتریس کاربر-گزینه که تاثیرات اجتماعی را درنظر گرفته است، درصورتیکه ما دریافته ایم که با استفاده از تاثیر اجتماعی روی شاخه مبتنی بر همکاری در سیستم هیبرید ما، به طور ضعیفی عمل کرده است. بعلاوه ما این سیستم را با ارزیابی ترتیب نسبی محلهای نزدیک در مقایسه با حضور واقعی به جای محاسبه صحت اماری درجه بندی های تخمین زده شده حین انجامشان ارزیابی کرده ایم. به عقیده ما، روش اجرای ارزیابی ما به رفتار پیشنهادات واقعی یک سیستم مبتنی بر موقعیت نزدیکتر می باشد. بنابراین، براین باوریم که کار ما تکمیل کننده کار ایشان در جنبه های مهم دیگر سیستم های پیشنهاددهنده اجتماعی مبتنی بر موقعیت کاربر می باشد که قبلا درنظر گرفته نشده بود.

تحلیل گرایشات و متن: در این کار، ما استفاده از برخی تکنیک های متن کاوی پیشرفته تر و قبلا گفته شده را برای ساخت یک سیستم پیشنهاد دهنده مبتنی بر مرور صرف مطرح کرده ایم. چندین تکنیک مدلسازی متن

تحت طرح پیشنهادات ما ارزیابی شده است مانند تخصیص هدایت پذیری نهفته LDA ، تحلیل نحوی نهفته LSA، یا TF-IDF . تحلیل گرایشات و احساسات همچنین در این مدلها با استفاده از طبقه بندی کننده های معمولی اموزش دیده نظیر محفظه های ساده Naïve Bayes، رگرسیون لوجستیک چندنامی، یا طبقه بندی کننده LDA نظارت شده لحاظ شده است.

تعداد بسیاری سیستم های پیشنهاددهنده دیگر وجود دارد که از اطلاعات یافت شده از متن دارای شکل آزاد حداکثر استفاده را می نمایند. در حقیقت، کار قبلی نشان داده است که درنظرگیری منابع متنی باعث بهبود تکنیک های فیلترسازی همکارانه استاندارد می شود. به عنوان مثال، Aciar و همکارانش یک سیستم توصیه کننده ای را برای محصولات مصرف کننده براساس مرورهای محصول ساخته سات یا Jakob و همکارانش پیشنهادات فیلم را براساس مرورهای فیلم بهبود داده است. Reschke و همکارانش یک سیستم گفتگوی پیشنهادات را مطرح کرده اند که براساس سوالات محدود از سوی مرورها ساخته شده است که تا اندازه ای متفاوت از تعریف مسئله پیشنهادات می باشد. برعکس همه این سیستم ها، سیستم پیشنهاددهنده ما، کل پیشنهادات را براساس داده های متنی مرورها بنا نهاده که به شیوه ای قابل انعطاف و کلی به جای ایجاد یک هستی شناسی متن یا استخراج جنبه های متعدد مدلسازی شده است.

زمینه اجتماعی برای بهبود نتایج در مسائلی سوای توصیه استفاده شده است. یک مثال که مستقیما به کار ما در اینجا مربوط می شود، مسئله تحلیل متن و بویژه شناسایی کیفیت مرور است. اغلب کارهای قبلی در این حوزه هر مروری را به شکل یک سند متنی مستقل فرض می کند و ویژگی هایی را از متن استخراج کرده و یک عملی را براساس ویژگی ها یاد می گیرد. به طور طبیعی، درنظرگیری اطلاعات دیگر به شکل روابط اجتماعی بین نویسندگان مرورها به بهبود پیشگویی ها کمک می کند.

پیشنهاددهنده های اجتماعی هیبرید: تعداد زیادی از محققان تکنیک های پیشنهادی متعددی را برای شکوفایی عملکرد سیستم های به اصطلاح هیبرید ترکیب کرده اند. سیستم های پیشنهادات مبتنی بر همکاری مستقل یا مبتنی بر محتوا از کمبودهای متعددی رنج می برند که می تواند با درنظرگیری درجه بندی های فردی شان غلبه شود. در این تحقیق، ما نشان داده ایم که این کمبودهای پیشنهاددهنده های انفرادی همچنین در شرح مشکل ما رخ می دهد و مزیت های مختلط سازی را روشن می سازد. در میان یک مجموعه وسیع از تکنیک های مختلط

سازی، ترکیب توزین شده خطی به نظر می رسد که یکی از ساده ترین بلکه موثرترین مکانیسم ها می باشد و یکی است که در اینجا اجرا می کنیم. برای مثال، Mobasher و همکارانش دو جز را که یک شاخه مبتنی بر همکاری و یک شاخه مبتنی بر محتوا می باشد به کمک یک طرح توزین خطی برای اجرای پیشنهادات فیلم ترکیب کرده بود. بخاطر مقایسه، ما باز دو مکانیسم هیبرید دیگر را تدوین کرده ایم به نام هیبرید ابشاری و هیبرید در سطح کلان که Burke دریافت به خوبی هنگام ترکیب دو مولفه به استقامت مختلف کار می کند. یک هیبرید ابشاری به نام Entree توسط Burke ساخته شد که یک پیشنهاددهنده مبتنی بر دانش و یک پیشنهاددهنده همکاری کننده را به حالت سلسله مراتبی براساس قدرت پیشنهادات شان ترکیب کرده بود. سرانجام اینکه یک هیبرید در سطح کلان که از پیشنهادات مبتنی بر محتوا برای شناسایی همسایگان همکاری کننده استفاده کرده است را می توان در رفرانس Pazzani در شماره ۳۸ یافت.

۳-فرایند بازیابی داده ها

۳-۱-شبکه های اجتماعی به عنوان منابع داده های باز

شبکه های اجتماعی (فیس بوک، توییتر، Foursquare، LinkedIn، Instagram، و غیره) به شکل انبارهای داده های جمع آوری شده و یکی شده از طریق متحدسازی فعالیت های اجتماعی کاربران مجزا به شکل طرحهای داده های مشترک عمل می کند، که اغلب بوسیله سیستم های تحلیلی یک سازمان یا گروه کاوش شده یا اینکه دردسترس جوامع تولیدکننده اپلیکیشن می باشد. این پلتفورم ها سیاستگزاری امنیت و محرمانگی داده ها را اجرا می کند، توسط کاربر چشم بسته پذیرفته شده که بر قابلیت دسترسی به داده های اجتماعی کاربران حکمرانی می کند.

ما به داده های باز اجتماعی به شکل محتوا اشاره می کنیم (پست های ارسالی، توییت ها، نکات مهم، انتشارات، عکس ها وغیره) که برای هرکسی مرئی یا عمومی می باشد. در دهه اخیر، داده های اجتماعی باز در بسیاری عرصه های تحقیقاتی از سیستم های پیشنهاددهنده برای طراحی و برنامه ریزی شهری ضروری شده اند. در میان همه انواع داده های باز اجتماعی، داده های موقعیت یابی جغرافیایی یک دامنه وسیعی از انواع رسانه ها را تحت پوشش قرار می دهند که شامل مختصات جغرافیایی می باشند. این امردر مورد شبکه های اجتماعی مبتنی بر موقعیت جغرافیایی مصداق دارد که برخی رسانه ها را (توییت، نکات مهم و غیره) به موقعیت کاربر که از سرتاسر سیستم

GPS گوشی هوشمند جمع اوری کرده است، مرتبط می سازد. همچنین شهرهای مدرن حاوی هزاران سنسور فیزیکی توزیع شده در نواحی جغرافیایی بزرگ را درنظر می گیرد که از آن اطلاعات حسگری شده آنها (یعنی الودگی هوا، ترافیک، سطح نور) به مختصات جغرافیایی مرتبط می شود.

داده های براساس موقعیت جغرافیایی که یا توسط کاربران اقدام کننده به عنوان سنسورهای اجتماعی یا توسط سنسورهای فیزیکی واقعی ایجاد شده است، می تواند از فراهم کنندگان داده های طرف ثالث (شبکه های اجتماعی، پورتالهای داده های باز، غیره) بوسیله تحقیقات فضایی زمینی مورد مشاوره قرار گیرد. معمولا، این هویت ها مکانیسمهای کنترل را برای اجتناب از بار بیش از حد ترافیک سرور اجرا می کنند که بازیابی تمامیت داده ها را فورا مختل می کند. با اینحساب، هر گونه فرایند بازیابی برای داده های تعیین موقعیت جغرافیایی باید به دقت این محدودیت های ترافیکی را درنظر بگیرد.

۲-۳-بازیابی داده ها از شبکه های اجتماعی

اغلب شبکه های اجتماعی به داده های خود از طریق اینترفیس برنامه ریزی اپلیکیسن (API) انتقال حالت نماینده REST دسترسی دارند. این پروتکل باعث یک تعامل اسان ولیکن موثر با محتوای داده های شبکه اجتماعی می گردد. ولیکن، این خدمات وب معمولا میزان تقاضاها را به ازای اپلیکیشن ثبت نامی و حجم داده های پاسخ برای اجتناب از بار اضافی ترافیک وارده محدود می سازد.

برای مثال، صف فضایی جغرافیایی برای کسب همه محل ها یا مکان های یک حیطه جغرافیایی از شبکه اجتماعی Foursquare برای برگشت ۵۰ محل حداکثر محدود شده است. این امر الزاما بیان می دارد که این ناحیه باید به زیرنواحی کوچکتری با کمتر از ۵۰ محل برای هر یک برای بازیابی موثر همه محل ها تقسیم بندی گردد. بعلاوه، پلتفورم Foursquare نیز تا ۵ هزار تقاضا در ساعت و در هر تقاضای ثبت نام شده محدود می شود و فرایند بازیابی را مجبور به اولویت بندی تقاضاهای بهینه می سازد. در اینجا ما استفاده از ساختارهای Quadtree و الگوریتم ساخت Quadtree را مطرح کرده و به انگیزه دراورده ایم تا بازیابی موثر و کارامد همه محتوای موقعیت جغرافیایی از یک ناحیه معنی ایجاد شده در شبکه های اجتماعی مبتنی بر موقعیت کاربر که محدودیت های فوق را نشان می دهند، انجام گیرد.

۳-۳-الگوریتم Quadtree

یک Quadtree یک ساختار داده ها است که براساس اصل تجزیه بازگشتی فضا می باشد. Quadtreeها به عنوان نمایشات داده سلسله مراتبی در حوزه های بینایی کامپیوتری، پردازش تصویری، تشخیص الگو و سیستم های اطلاعات جغرافیایی استفاده می شود. علاقه به این ساختار داده ها ریشه در این حقیقت دارد که برای تمرکز منابع روی حیطه هایی طراحی شده که در آن اطلاعات از بیشترین تراکم برخوردار است. برای یک تحقیق جامع روی Quadtreeها و ساختارهای سلسله مراتبی مربوطه به رفرانس Samet شماره ۴۲ مراجعه کنید.

طرح ساخت Quadtree در الگوریتم ۱ ارائه شده است و به ترتیب ذیل کار می کند. این الگوریتم به طور تکراری هر ناحیه یا سلول را به چهار زیرناحیه یا زیرسلول تقسیم بندی می کند وقتی که ظرفیت ماکزیمم هر سلول که با N_{max} نمایش داده می شود به حدنصاب برسد. با درنظرگیری دو مختصات (محدوده های جنوب غربی و شمال شرقی) که تعریف کننده کادر محدودکننده جغرافیایی می باشد، این الگوریتم یک شی Quadcell را تعریف کرده و به صف می کند که حاوی حدود جغرافیایی اش می باشد.

الگوریتم ۱-الگوریتم Quadtree

```
quad ← Quadcell(NElim, SWlim);
queue ← List(quad);
while Length(queue) > 0 do
    q ← Pop(queue;          //    Obtains quad from the queue
    if CheckQuad(q) then
        aux ← SplitQuad(q);
        Extend(queue, aux);          //    Push quad into the queue
    end
end
```

برای هر quadcell در صف، این الگوریتم از شبکه اجتماعی REST API استفهام می کند و اهمیت پاسخ را طبق تعداد ماکزیمم سنسورها در هر سلول N_{max} مقایسه می کند. این کار با تابع CheckQuad توضیح داده شده در الگوریتم ۲ اجرا می شود.

الگوریتم ۲-تابع Quadcell بررسی

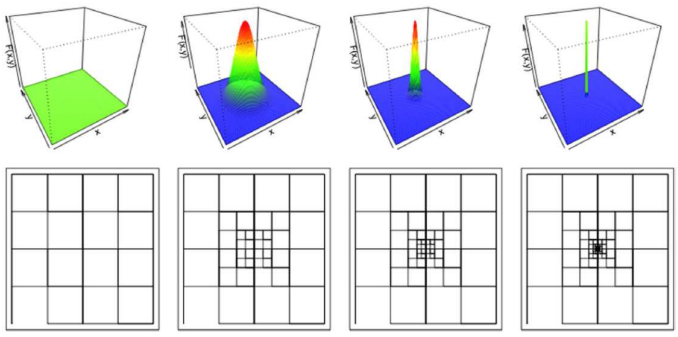
```
Function CheckQuad(quad: Quadcell) : boolean
    sensors ← APIrequest(quad.NElim, quad.SWlim);
    if Length(sensors) < N_max then
        SaveSensors(sensors);
        return False;          //    Do not split quad
    else
        return True            //    Splitquad
    end
end
```

در حالتی که اندازه پاسخ از N_{max} تجاوز کند، Quadcell به چهار زیرناحیه یا فرزند تقسیم بندی می شود، که محدودیت های جغرافیایی اش از روی مرزهای والدین آن محاسبه شده و اینها همچنین به صف معوق اضافه می شوند. این کار با تابع SplitQuad که در الگوریتم ۳ توضیح داده شده است، اجرا می شود. ترک ها یا Quadcellهایی که استفهام API آن سنسورهای کمتر از ماکزیمم مجاز را بازگردانده است، به شکل دیسک ذخیره سازی شده و Quadcellها از صف حذف می شوند.

الگوریتم ۳-تابع Quadcell جداسازی شده

```
Function SplitQuad(quad: Quadcell) : List
        //    Computes the NElim and SWlim for each child
    NElims, SElims, SWlims, NWlims ← ChildrenLim(quad.NElim, quad.SWlim);

        //    Creates Listof children
    children ← List(Quadcell(NElims.NE, NElims.SW),
    Quadcell(SElims.NE, SElims.SW),
    Quadcell(SWlims.NE, SWlims.SW),
    Quadcell(NWlims.NE, NWlims.SW);
    return children;
end
```



شکل ۲-در بالا، توزیع فضایی داده ها با موقعیت جغرافیایی یکنواخت بوده و به طور عادی با انحراف استاندارد $\sigma = 100, 10, 0.25$ توزیع می شود. ساختارهای Quadtree متناظر در پایین.

برای تفسیر نتایج الگوریتم Quadtree برای سنسورهایی که روی یک ناحیه محدود پخش شده اند، ما اجرای Quadtree را برای توزیعات فضایی مختلفی شبیه سازی کرده ایم که همگی حاوی تعداد برابر سنسورها می باشند (بالای شکل ۲). به عبارت دیگر، همه توزیعات به N جمع بندی شده که تعداد سنسورها در ناحیه می باشد. ساختارهای Quadtree نتیجه شده (پایین شکل ۲) به وضوح نشان دهنده وابستگی میان اوج توزیع و تعداد Quadcell ها می باشد. با درنظرگیری اینکه هر quadcell دست کم یک استفهام HTTP را برای REST API بیان می کند، توزیعات اوج داده های فضای جغرافیایی منجر به اجراهای quadtree با کارایی کمتر در مقایسه با توزیعات یکنواخت تر و ترازتر می شود.

الگوریتم Quadtree بر محدودیت های API با اولین تقسیم بندی فضای جغرافیایی به زیرنواحی مناسب و بعد با محدود نگه داشتن میزان تقاضای HTTP غلبه می کند. ما دستورالعمل های مناسبی را برای فرایند کنترل روی میزان تقاضا برای ساده نگه داشتن توضیح در الگوریتم ۱ اضافه نکرده ایم. ولیکن اجرای الگوریتم Quadtree باید کنترل را طی زمان بین تقاضاهای پی در پی یا تعداد کلی تقاضاها در هر ساعت حفظ کند و برنامه هنگام لزوم غیرفعال شود.

۴-۳-مقیاس بندی الگوریتم Quadtree

کارایی و قابلیت مقیاس بندی با حجم امروزی داده های در دسترس یک اصل مهم به حساب می اید. در اینجا، ما یک نسخه موازی از الگوریتم Quadtree را مطرح کرده ایم. ابتدا ما موانع قابلیت مقیاس بندی اصلی الگوریتم مطرح شده را بیان کرده ایم و بعد از الگوریتم بازدید مجدد کرده ایم تا یک الگوریتم موازی جدید Quadtree را مطرح سازیم.

ما فرض کرده ایم که ماکزیمم تعداد تقاضا در ساعت و در هر تقاضای ثبت شده برابر با R_{max} می باشد. این محدودیت توسط مجری شبکه اجتماعی اعمال شده است. وقتی تعداد سنسورهای N بزرگ باشد مقدار R_{max} عامل محدودکننده عملکرد الگوریتم می شود.

شیوه ما از موازی سازی الگوریتم به k زیرفرایند استفاده کرده است، که هر یک از یک کلید اپلیکیشن ثبت نام شده متفاوتی استفاده کرده و با اینحساب باعث میزان تقاضای بیشتری می باشد (با K زیرفرایند که حد آن

KR_{max} می باشد). این زیرفرایندها یا در همان ماشین یا در ماشین های مختلفی با نشانی IP عمومی مجزایی بسته به رهنمودهای شبکه اجتماعی اجرا می شود.

یک شیوه کلاسیک برای رهگیری این مسئله چندمرحله ای بوسیله پارادیگم تولیدکننده-مصرف کننده است. تولیدکننده مسئول سوال از شبکه اجتماعی API ، ایجاد و ردیف سازی quadcellها به یک ردیف پردازشی می باشد. بنابراین، مصرف کننده هر quadcell ردیف شده را گرفته و انرا به شکل دیسک مرتب سازی می کند. با تقسیم بندی دو زیرفرایند با بیشترین زمانبری (مرتب سازی API و مرتب سازی به شکل دیسک)، ما بازده قابل ملاحظه ای را به موازات الگوریتم Quadtree کسب کرده ایم.

طرح تولیدکننده-مصرف کننده برای موازی سازی Quadtree که در الگوریتم ۴ و ۵ توضیح داده شده است به ترتیب ذیل کار می کند. رشته های تولیدکننده (الگوریتم ۴) یک سری Quadcell هایی (prodQuad) را با مرتب سازی شبکه اجتماعی API درباره quadcellهای مرتب سازی شده در ردیف معلق (PendQueu) تولید می کند، و بررسی می کند که آیا اهمیت پاسخ ها از تعداد ماکزیمم سنسورها در هر خانه (N_{max}) افزایش یافته است یا خیر و نیز quadcellها را اگر زیاد شده باشد تقسیم بندی می کند. Quadcellهای تقسیم که تعداد سنسورهای آن از N_{max} بیشتر شده باشد همچنین به صف معوق برای تولیدکننده بعدی ارسال می شود. بعد، تولیدکننده ترک ها یا quadcellهایی را که سنسورهای کمتر N_{max} دارند، به شکل ردیف پردازشی (ProcQueue) ذخیره سازی می کند، و به متغیر شرایط (Cond) برای رهایی قفل زیربنایی اخطار می دهد. این اخطار رشته های مصرف کننده (الگوریتم ۵) را بیدار می سازد که منتظر متغیر شرایط می باشند که quadcellها (consQuad) را از صف پردازشی مصرف می کند. مصرف به معنای کشاندن quadcell از صف و ذخیره سازی مقادیر سنسورها به شکل دیسک می باشد. وقتی که رشته مصرف کننده کار خودش را تکمیل کرد، منتظر متغیرشرایط برای رهایی بعدی در اثر تولیدکننده ها می شود.

الگوریتم ۴-الگوریتم Quadtree تولیدکننده

```
Function Producer (PendQueue: Queue, ProcQueue: Queue,Cond: Condition,API: APIhand) : void
  while countQuads(PendQueue) > 0 do
    with(Cond)
      q = prodQuad(PendQueue, API);
      putQuad(ProcQueue, Q);
      notify(Cond);
  end
  with Cond       ::
    notifyAll(Cond)
end
```

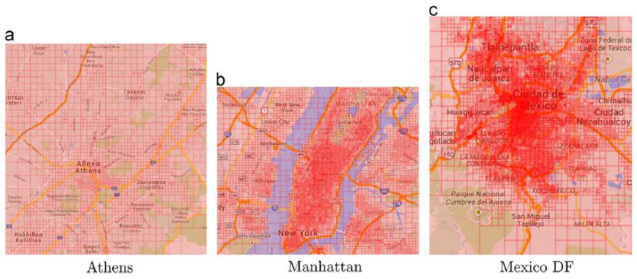
الگوریتم ۵-الگوریتم Quadtree مصرف کننده

```
Function Consumer (PendQueue: Queue, ProcQueue: Queue, Cond: Condition) : void
  while countQuads(PendQueue) > 0 do
      with(Cond);
      Wait(Cond);
      if countQuads(ProcQueue) > 0 then
        consQuad(ProcQueue)
      end
  end
end
```

۵-۳-مطالعه موردی: پلتفورم Foursquare

پلتفورم Foursquare یک REST API را فراهم می کند تا با محتوای آن در تعامل باشد و نیز به داده های اجتماعی باز آن از طریق اپلیکیشن های ثبت شده دسترسی یابد. استفهام از پلتفورم درباره داده های محلهای با موقعیت جغرافیایی مستلزم رعایت محدودیت های اندازه پاسخ (۵۰ محل به ازای هر تقاضا) و محدودیت های درجه بندی (۵ هزار تقاضا در هر ساعت و تقاضا) می باشد. الگوریتم موازی Quadtree باعث یک فرایند بازیابی داده های موثر و کارامد با موازی سازی تقاضاهای API به زیرفرایند K_p و مرتب سازی برای دیسک سازی زیرفرایندهای K_c می شود. ما دریافته ایم که استفاده از $K_p=3$ تولیدکننده و $K_c=10$ مصرف کننده یک عملکرد مناسبی را نشان میدهد هر چند بهینه سازی این پارامترها فراتر از دیدگاه این مقاله است.

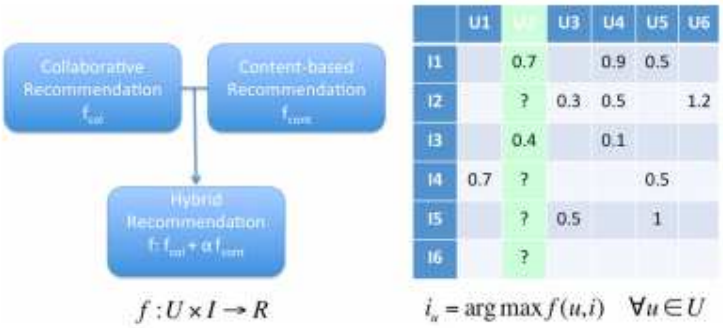
شکل ۳ نشان دهنده ساختارهای Quadtree برای سه ناحیه شهری بزرگ است: شهر آتن، منهتن، و مکزیکوسیتی. همانگونه که می توانیم انتظار داشته باشیم، Quadtree در پایین شهر و نواحی فرعی تجاری و بازرگانی بیشتر رشد کرده درحالیکه در نواحی مسکونی، پارکها و حومه شهر کمتر رشد دارد. جدول ۱ نشاندهنده برخی ویژگی ها درباره Quadtrees است که در هر یک از این نواحی شهری ساخته شده است.



شکل ۳-ساختارهای Quadtree موازی. a) آتن، b) منهتن، و c) مکزیکو DF.

جدول ۱-Quadtree شهری دارای $K_p=3$ تولیدکننده و $K_c=10$ مصرف کننده

Features	*Athens*	*New York*	*Mexico D.F.*
NE lim	38.03, 23.79	40.80, −73.91	19.59, −98.94
SW lim	37.95, 23.69	40.70, −74.07	19.13, −99.36
Venues	48.215	297.924	511.096
Quadcells	2.997	18.682	32.270
Quadcell leaves	2.248	14.011	24.202
Quadtree time (s)	4.432	27.980	32.015



شکل ۴-یک مرور اجمالی گرافیکی سیستم پیشنهاددهنده هیبرید و ماتریس پیشنهادات آن

۴-شرح سیستم GeoSRS

۴-۱-مرور اجمالی

سیستم GeoSRS یک طرح مبتنی بر درجه بندی برای اجرای پیشنهاد گزینه می باشد. این بدان معناست که بادرنظرگیری یک تابع استفاده $f: U \times I \rightarrow R$ ، درجه یا سنجش اهمیت از روی مجموعه کلیه کاربران U و مجموعه کلیه گزینه های I تخمین زده می شود. مسئله پیشنهاد را می توان به شکل مسئله بهینه سازی ذیل دید:

$$i_u = \arg\max_{i \in I} f(u,i) \quad \forall u \in U \qquad (1)$$

که در آن برای هر کاربر u، هدف تعیین گزینه i است که تابع استفاده را در میان همه گزینه های موجود به حداکثر می رساند. گزینه i_u که این به حداکثررسانی را رعایت می کند به کاربر u پیشنهاد می شود. ولیکن، تابع استفاده معمولا برای کل فضای $U \times I$ تعریف نشده است، و در نتیجه، پیشنهاد کننده نیاز دارد که درجه بندی از دست رفته را برای اجرای یک پیشنهاد صحیح تخمین بزند همانگونه که در ماتریس پیشنهاد از روی شکل ۴ نشان داده شده است.

GeoSRS یک سیستم پیشنهاددهنده اجتماعی هیبرید است که یک سیستم فرعی مبتنی بر همکاری و مبتنی بر محتوا را ترکیب کرده تا تابع استفاده را تعریف کند که درجه بندی های گزینه های دیده نشده را تخمین می زند. بویژه، تابع استفاده از روی میل ترکیبی کاربر-گزینه و نیز عقیده اجتماع با ترکیب انها به شکل یک سیستم هیبرید، با در نظرگیری ترکیب خطی ذیل تعریف می شود:

$$f = f_{col} + \alpha f_{cont}$$

که در آن α یک مقدار واقعی مثبت است و در حالیکه شاخه مبتنی بر محتوای f_{cont} اساسا از ویژگی های استنباط شده گزینه و مشخصات کاربر استفاده کرده، شاخه جمعی f_{col} از مزیت خرد جمع بهره می گیرد.

۲-۴-شاخه مبتنی بر محتوا

شاخه پیشنهاد مبتنی بر محتوا به سیستم ترجیحات کاربر و ویژگی های گزینه را با هدف انجام پیشنهاد هایی طبق مشخصات کاربر می آورد. بدنبال نامگذاری معرفی شده، تابع فایده برای شاخه مبتنی بر محتوا میزان گزینه های دیده شده و دیده نشده را پیشگویی می کند. میزان R از روی محصول نقطه ای بردارهای مشخصات کاربران w_u و بردارهای ویژگی گزینه های w_i پیشگویی می شود. با اینحساب، تابع فایده می تواند به شکل ذیل مجددا تعریف گردد:

$$f_{cont}: w_u \times w_i \rightarrow R \qquad (2)$$

به دلیل این حقیقت که فضاهای گزینه ها و کاربران بزرگ می باشند، مکانیسم کاهش بعدگرایی یا فیلترسازی گزینه ها/کاربران اغلب استفاده می شود.

۱-۲-۴-اندازه گیری شباهت

اندازه گیری شباهت در میان جفت کاربران و گزینه ها می تواند یا با استفاده از مدلهای کاوشی یا مدلهای آماری یادگیری شده از روی داده های زیربنایی محاسبه گردد. شیوه مطرح شده ما برای شاخه مبتنی بر محتوا از یک مدل کاوشی مبتنی بر کسینوس برای محاسبه میزان شباهت بین مشخصات کاربر و ویژگی های گزینه ها استفاده می کند. با اینحساب، فرمولاسیون تابع فایده می تواند به شکل ذیل بیان شود:

$$f_{cont} = \cos(w_u, w_i) = \frac{\overrightarrow{w_u}\,\overrightarrow{w_i}}{\| \overrightarrow{w_u} \| \, \| \overrightarrow{w_i} \|} \qquad (3)$$

که در آن w_u و w_i به ترتیب بردارهای مشخصات برای کاربران و گزینه هایی هستند که از روی مدلهای محتوایی مرورها که در ذیل توضیح داده شده، تعریف شده اند.

۲-۲-۴-مدلهای محتوایی مرور

مرورهای متنی با استفاده از تکنیک های استاندارد پیش پردازش شده اند، بویژه ما کلمات توقف و علائم نوشتاری، فواصل سفید خالی را برداشته ایم و همه متن را به شکل حروف کوچک تبدیل کرده ایم.

در این مقاله ما چندین تکنیک پیشرفته را برای مدلسازی مرور درنظر گرفته ایم. هدف از مدلسازی مرورها ساخت مشخصات توصیفی برای هم کاربران و هم واژه ها می باشد. تا به اینجا، ما همه مرورهای یک کاربر را جمع اوری کرده ایم، و مرورهای جمع اوری شده را یک سند متنی منفرد درنظر گرفته که مدلسازی به شکل w_u نماییم. به طور مشابهی، ما همه مرورهای یک گزینه معین را برای ساخت یک مدرک جمع اوری کرده ایم که فورا مدلسازی شده اند و یک گزینه w_i را مشخص خواهند کرد. درآنچه بدنبال می آید، ما تکنیک های مدلسازی را که استفاده کرده ایم برای مدلسازی مرورهای متنی جمع اوری شده به شکل K مولفه برای هر دو بردار، فهرست کرده و به اختصار توضیح می دهیم.

فراوانی واژه- فراوانی مدرک معکوس: TF-IDF محتوای یک مرور را به یک مجموعه از K کلیدواژه بسته به اندازه گیری اهمیت از طریق محاسبه یک آمار عددی متناظر می سازد. آمار عددی برای هر کلمه از مرور محاسبه شده و بعد K کلمه مرتبط تر نگه داشته می شود.

تحلیل معنایی نهفته: LSA می تواند برای متناظرسازی محتوای یک مرور به K مفاهیم نهفته متناظر شود که مشخص می شود کلماتی هستند که معنای نزدیکی دارند. LSA با بکارگیری تجزیه ارزش منفرد یک ماتریس واژه-مدرک اجرا می شود. این ماتریس حاوی واژه ها یا کلماتی در ردیف ها و اسناد یا مرورها در ستونها می باشد. عدد واقعی در تقاطع ردیف ها و ستونها نشان دهنده رخداد واژه ها درون متن است. k بزرگترین ارزشهای ویژه و بردارهای ویژه متناظر آنها منجر به k تخمین رتبه ای ماتریس واژه-سند می شود.

تخصیص هدایت پذیری نهفته: LDA یک مدل احتمال گرایانه عمومی است که در آن اسناد یا مرورها به شکل مخلوطهای تصادفی روی موضوعات نهفته ارائه می شود، و هر موضوع با یک توزیع روی کلمات مشخصه سازی شده است. LDA برای مدلسازی محتوای مرورها در زمینه موضوعاتی استفاده می شود که در یک مرور رخ می دهد. به لحاظ شهودی، یک مرور می تواند درباره خدمات یک رستوران ژاپنی صحبت کند و مرور دیگر به غذای یک رستوران محلی اشاره کند. هر دو مرور منجر به موضوعات مختلفی می شود اگر فضای موضوع به قدر کافی بزرگ باشد که در میان آنها تمایز قائل شود.

۳-۴-شاخه همکاری

شاخه همکاری سیستم پیشنهاددهنده مبتنی بر مرور با هدف جمع آوری عقاید جمع از مرورهای ارسالی کار می کند. پیشنهاد براساس این عقیده می باشدکه محله نزدیک درباره یک گزینه خاص می باشد. با اینحساب ما تابع فایده f را برای گزینه های دیده شده $r_{u',i}$ جوری تعریف می کنیم که شاخص گرایشات و احساسات از تحلیل گرایشات و احساسات یک مرور متنی باشد که یک کاربر درباره یک گزینه نوشته است. برای گزینه های دیده نشده، کاوش یک میانگین گیری را روی درجات آندسته کاربران مشابه (همسایگان) را انجام می دهد که گزینه را دیده باشند. به طور فرمولی:

$$f_{col} = r_{u,i} = \frac{1}{N} \sum_{u' \in N_u} r_{u',i} \qquad (4)$$

که در آن N_u نمایانگر N تعداد کاربران در محله u می باشند. تعیین N و محله کلیدی برای دستیابی به میزان خوب تخمین برای گزینه های نادیده می باشد.

۱-۳-۴-شناسایی محله

شناسایی محله یک نقش مهمی را برای دستیابی به یک پیشگویی صحیح درجه پیشنهاددهنده بازی می کند. سیستم های پیشنهاددهنده به طور معمول از شباهت در میان کاربران از ماتریس درجه بندی کاربر-گزینه استفاده کرده اند. شباهت معمولا جفت های درجه بندی هایی را به حساب می آورد که هر دو کاربران درجه بندی کرده اند. اندازه گیری شباهت معمولا بوسیله یک مکانیسم کاوشی مانند ضریب همبستگی پیرسون یا شباهت مبتنی بر کسینوس محاسبه شده است.

ما استفاده از شباهت مبتنی بر کسینوس را برای شاخه همکاری پیشنهاددهنده مبتنی بر مرور مطرح کرده ایم که به شکل اندازه گیری کاوشی شباهت دو جفت بردارهای کاربر $\overrightarrow{u_x}, \overrightarrow{u_y}$ می باشد. یعنی:

$$sim(u_x, u_y) = \cos(u_x, u_y) = \frac{\overrightarrow{u_x}\overrightarrow{u_y}}{\|\overrightarrow{u_x}\| \|\overrightarrow{u_y}\|} \quad (5)$$

بردارهای کاربر از درجه بندی های گرایشات طی همه گزینه ها تعریف می شود. با این شیوه، جفت های کاربران، که هم جفت های گزینه ها را مثبت و هم منفی درجه بندی کرده اند، عقاید مشابهی را برای گزینه های دیده نشده به اشتراک می گذارند. شباهت با پایه کسینوسی سنجش مفیدی برای شناسایی آنها می شود.

اخیرا، سیستم های پیشنهاد دهنده اجتماعی مکانیسم تازه ای را برای شناسایی محله برای فیلترسازی جمعی ارائه داده اند. Groh و Ehmig استفاده از دوستی اجتماعی یا روابط را برای ایجاد محله برای یک پیشنهاددهنده مبتنی بر همکاری مطرح کرده اند. طبق گفته آنها، این شیوه می تواند از فیلترسازی جمعی معمولی بهتر عمل کند.

۲-۳-۴-مدلهای گرایشات مرور

مرورهای متنی همچنین با استفاده از تکنیک های قبلی قبل از بکارگیری مدلهای گرایشات نظارت شده ذیل پیش پردازش می شوند. این مدلها با داده های مرورهای برچسب دار شده اموزش دیده و برای پیشگویی گرایشات برای بررسی های خارج از نمونه استفاده می شوند. در واقع، ما از سه مجموعه داده های مختلف برای ساخت، تست و روایی سازی طبقه بندی کننده های گرایشات استفاده کرده ایم:

-یک مجموعه داده های آموزشی: این مجموعه متشکل از ۱۵۰۰ مرور متنی است. این مرورهای متنی با استفاده از AFINN نشانگذاری شده است، که فهرستی از کلمات انگلیسی است که برای ظرفیت با یک عدد صحیح بین

منهای پنج (منفی) و مثبت پنج (مثبت) درجه بندی شده است. درجه بندی به سه مقدار ۱- و ۰ و ۱ از طریق یک نسخه اصلاح شده تابع علامت گذاری مجزا می شود که نشان دهنده مرورهای به ترتیب منفی، خنثی و مثبت می باشد.

-یک مجموعه داده های تست: مرکب از ۵۰۰ مرور متنی است که همچنین با استفاده از روش AFINN نشانگذاری شده است.

-یک مجموعه داده هایی روایی: مرکب از ۲۰۰ مرور متنی است که به طور دستی به صورت مرورهای مثبت، خنثی یا منفی نشانگذاری کرده ایم.

طبقه بندی کننده گرایشات محفظه تازه برنولی: یک طبقه بندی کننده محفظه های ساده یک مدل احتمال گرایانه را با فرضیات مستقل ساده بین ویژگی هایی برای پیشگویی توزیع احتمالات یک نمونه روی مجموعه طبقات درنظر می گیرد. ما با هدف مدلسازی ساده گرایشات مرور با استفاده از یک طبقه بندی کننده محفظه های ساده کار می کنیم که یک مدل کیسه ای از کلمات را برای متن مرور درنظر گرفته ایم. این امر بدان معناست که ما کلمات را به شکل ویژگی های مدل بدون درنظر گرفتن موقعیت آنها در مرور و فرض استقلال در میان کلمات یک رده خاص (مثبت، منفی یا خنثی) درنظر گرفته ایم. این امر می تواند با استفاده از مدل محفظه های ساده برنولی حاصل آید که ویژگی ها را به شکل ورودی های باینری مستقل تعریف می کند که توصیف می کند که آیا یا خیر یک کلمه در یک مرور معین وجود دارد یا خیر (به رفرانس ۳۱ برای یک طبقه بندی درباره انواع مختلف طبقه بندی کننده های محفظه های ساده مراجعه شود).

طبقه بندی کننده گرایشات رگرسیون لوجستیک چندنامی: چون هدف ما تخمین سه نتیجه احتمالی برای هر مرور میباشد، مرورهای مثبت، منفی و خنثی، یک طبقه بندی کننده چندرده ای مورد نیاز است. رگرسیون لوجستیک چندنامی رگرسیون لوجستیک را به یک تنظیم چندرده ای تعمیم می دهد. این طبقه بندی کننده احتمالات نتایج احتمالی را براساس یک مجموعه از ویژگی ها تخمین می زند. این طبقه بندی کننده از طبقه بندی کننده محفظه های ساده متفاوت است از این لحاظ که نیازی به استقلال آماری طبق مجموعه ای از ویژگی ها یا کلمات استفاده شده در رگرسیون منطقی چندنامی وجود ندارد. یک مانع در مقایسه با محفظه های ساده

این حقیقت می باشد که تخمین ضرایب همبستگی رگرسیون از یک طبقه بندی کننده چندنامی خیلی پیچیده تر است و عموما به یک فرایند تکراری نیازمند است.

تخصیص هدایت پذیری نهفته نظارت شده: طبقه بندی کننده SLDA برای تحلیل گرایشات باعث توسعه مدل موضوعی LDA با متغیرهای پاسخ دهنده برای هر سند می شود. متغیر پاسخ برای یک طبقه بندی کننده تحلیل گرایشات مربوط به رده گرایشات می شود: مثبت، منفی یا خنثی. اسناد و پاسخ های آنها به طور مشترک مدلسازی شده تا موضوعات نهفته را بیابد که به بهترین نحوی پاسخ هایی را برای اسناد بدون برچسب آتی پیشگویی کرده است.

۴-۴-تنظیمات هیبرید

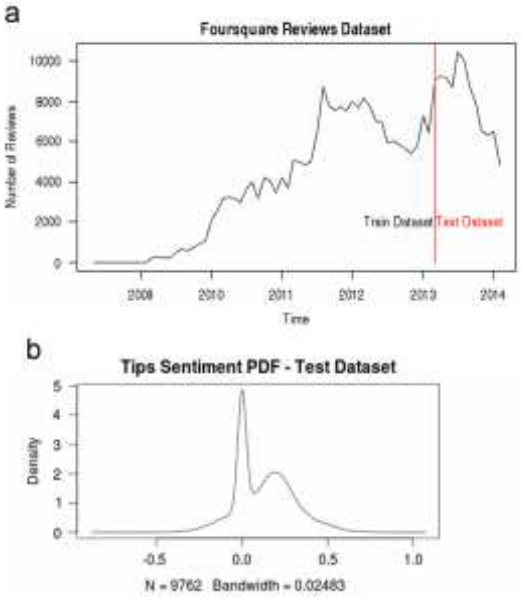
با استفاده از یک ترکیب خطی هیبرید از شاخه های همکارانه و مبتنی بر محتوا، ما وانمود می کنیم که موانع هر شیوه انفرادی را به حداقل رسانیده ایم. مدل هیبرید می تواند به ترتیب ذیل با $\alpha > 0$ نمایان شود،

$$f: U \times I \to R \Rightarrow f: f_{col} + \alpha f_{cont} \quad (6)$$

از یک سو، شیوه های همکاری تا حد زیادی به وجود توده ای از کاربران منتقد بستگی دارد که گزینه های کافی را برای پیشگویی موثر گزینه های نادیده درجه بندی کرده باشند. یک شیوه برای غلبه بر نادر بودن ماتریس درجه بندی کاربر-گزینه استفاده مشترک از داده های نمایه سازی شده برای شناسایی محله نزدیک می باشد. روش ما به حداقل رسانی مسئله نایاب بودن با ترکیب درجه بندی ها از شاخه مبتنی بر متن می باشد انهم زمانی که درجه بندی های همکارانه به قدر کافی صحیح نباشند.

از سوی دیگر، روشهای مبتنی بر محتوا اساسا در توصیه گزینه هایی شکست خورده اند که ویژگی هایشان تا کنون توسط کاربران درجه بندی نشده است. بعبارت دیگر، پیشنهاد دهنده مبتنی بر محتوا نسبت به ویژگی ها از گزینه های دیده شده بیش از حد تخصص یافته است چون نمایه سازی کاربر ناشی از گزینه هایی است که قبلا توسط آن کاربر دیده شده است. تنوع گزینه به سیستم ما اضافه می شود تا بر این تخصص بیش از حد غلبه نماید. با ترکیب گرایشات جمعی به سیستم مطرح شده، ویژگی های گزینه دیده نشده همچنین با تقویت صحت کل پیشنهاد می شود.

همچنین موانعی وجود دارد که مشترک هر دو شیوه می باشد نظیر مسئله شروع سرما. برای مثال، مسئله کاربر جدید زمانی رخ مید هد که یک کاربر هر گزینه یا معدودی را مرور نکرده باشد. در این موقعیت، پیشنهاددهنده قادر به اجرای پیشنهاد نیست و این امر پوشش سیستم را پایین می اورد. مسئله گزینه جدید نیز یک مانع است که حداقل سازی مشکلی در هر دو تنظیمات دارد. این امر زمانی رخ می دهد که یک گزینه برای سیستم جدید باشد و با اینحال توسط هر کاربری مرور و بررسی نشده باشد. چون هیچ فیدبک کاربری برای این گزینه ها وجود ندارد، این سیستم قادر نیست که یک پیشنهاد صحیح را اجرا نماید، که همچنین بر پوشش سیستم اثر می گذارد. این دو مانع اساسا به دلیل فقدان داده هاست و با اینحساب راه حل آنها الزاما متحمل فرایند خزش سازی موثر داده های بیشتری از سوی کاربران و گزینه ها می شود.



شکل ۵-توضیح مجموعه داده های Foursquare در ارزیابی ما. a) تعداد مرورهای رستوران در Foursquare طی زمان و b) تابع دانسیته احتمالات گرایشات نکات مهم

۵-ارزیابی سیستم

روش ارزیابی اندازه گیری خواهد کرد که به چه موثری سیستم GeoSRS یک گزینه دیده نشده را به یک کاربر معین پیشنهاد می دهد. سپس از شبکه اجتماعی برای ارزیابی صحت عملکرد سیستم با مقایسه فیدبک کاربر واقعی درباره یک گزینه علیه میزان توصیه احتمالی برای آن گزینه استفاده خواهد کرد. با اینحساب، ما فرض می

کنیم که یک درجه ای از علیت بین حقیقت خرید/تجربه یک گزینه و عمل بعدی مرور آن محصول/تجربه وجود دارد.

این شیوه همچنین یک مکانیسم ساده ولیکن قدرتمندی را برای اندازه گیری پوشش پیشنهادات یا صرفا نسبت پیشنهاداتی که سیستم قادر است حاصل اورد، فراهم کرده است. عدم توانایی انجام یک پیشنهاد می تواند به دلیل این حقایق رخ بدهد که یا کاربر فیدبکی را برای اولین بار فراهم کرده است (مسئله کاربر جدید) یا اینکه این گزینه تا کنون مرور نشده است (مسئله گزینه جدید). برای مثال، یک کاربری که در Foursquare یک محلی مانند یک رستوران را بررسی می کند که در آنجا غذایی خورده احتمالا فیدبک خودش را به شکل یک مروری در شبکه اجتماعی فراهم می سازد. تفسیر این فیدبک (مثبت، منفی یا خنثی) در روش ارزیابی ما استفاده می شود تا طبق میزان پیشنهادات از GeoSRS در زمان بررسی و درون محله آن رستوران مقایسه بشود.

بعد، ما مجموعه داده های Foursquare را توضیح می دهیم که بعدها برای ارزیابی سیستم GeoSRS از لحاظ پوشش پیشنهادات و صحت عملکرد استفاده می شود.

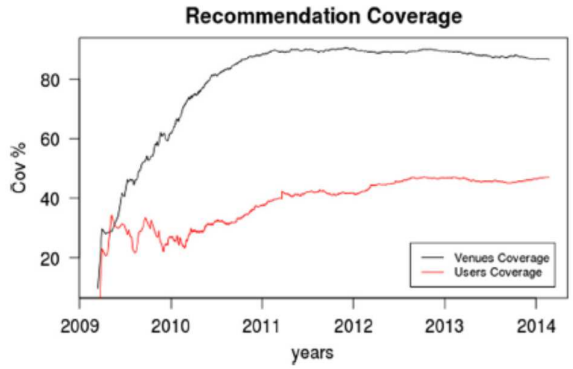
۱-۵-مجموعه داده های نکات مهم رستوران Foursquare

مجموعه داده های نکات مهم رستوران Foursquare شامل ۳۰۹,۶۴۰ مرور کوتاه یا نکات مهم از منطقه منهتن می باشد. کل مجموعه داده ها به مجموعه های آموزشی (۷۰ درصد از قدیمی ترین نکات مهم) و مجموعه های تست (۳۰ درصد از جدیدترین نکات مهم) تقسیم بندی می شود.

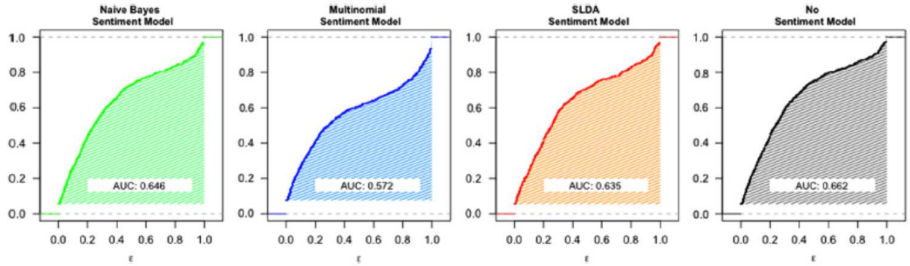
در شکل ۵a ، ما تعداد مرورهایی را به شکل تابعی از زمان نشان داده ایم، که از اولین مرور در سال ۲۰۰۸ تا اخرین اجرای خزش کاری در فوریه ۲۰۱۴ می باشد. بعلاوه، این گرافیک نشان دهنده یک تمایل مثبت در تعداد بررسی های ایجاد شده از سال ۲۰۰۸ می باشد که به ما نشان می دهد که Foursquare از آن زمان، علی رغم برخی الگوهای فصلی در حال رشد بوده است. همانگونه که گفته شده است، کل مجموعه داده ها به دو زیرمجموعه تقسیم بندی می شود که همچنین در نمودار نشان داده شده است.

شکل Fig. ۵b نشان دهنده توزیع قطبیت گرایشات نکات مهم در مجموعه داده های تست می باشد. درحالیکه مقادیر منفی در pdf بیان گننده گرایشات منفی است، گرایشات مثبت در حدفاصله [۰،۱) می گنجد. از سوی دیگر، فیدبک کاربر خنثی با یک مقدار گرایشات صفر نمایش داده می شود. برای اهداف ارزیابی، ما فیدبک خنثی

را به عنوان یک تجربه مثبت درنظر می گیریم چون بیشتر انها جملاتی نظیر «شکلات داغ را امتحان کن!» می باشد. ما این فرضیه را صورت داده ایم که کاربر به طور تلویحی تجربه را به شکل مثبت درجه بندی می کند. همچنین گفته ایم که مجموعه داده های تست از لحاظ درجه بندی های مثبت ومنفی نامتعادل است و واقعیت داده های نکات مهم Fuorsquare را از اوایل ۲۰۱۳ الی ۲۰۱۴ نشان می دهد. این امر ممکن است هنگام ارزیابی عملکرد صحت این مجموعه داده ها خیلی نامرتبط باشد، ولیکن هنگام مقایسه با تنظیمات پیشنهاددهنده مختلف و قابلیت های تعمیم مرتبط می شود.



شکل ۶-پوشش پیشنهادات برای محل ها و کاربران



شکل ۷-AUCها از cdfی خطای پیشنهادات برای انواع مدلهای گرایشات.

۲-۵-پوشش پیشنهادات

پوشش پیشنهادات حوزه گزینه ها را اندازه گیری می کند که روی آن سیستم می تواند پیشنهاداتی را اجرا نماید. معمولا، واژه پوشش هرماه با ۱)درصد گزینه هایی است که برای آنها سیستم قادر به تولید پیشنهاد است و ۲) درصدگزینه های موجود که به طور موثری پیشنهاد داده می شود. در اینجا، ما تعریف قبلی را اتخاذ کرده ایم چرا که ما دومی را در مفهوم صحت عملکرد تعریف شده بعدی اتخاذ کرده ایم. با اینحساب، به طور رسمی، پوشش را

به شکل درصد گزینه های موجود (I)تعریف می کنیم که برایش سیستم پیشنهاد دهنده می تواند یک پیشگویی (I_P) را حاصل نماید یعنی $Cov = (|I_P|/|I|) \cdot 100$ می باشد.

در سیستم GeoSRS و درون طرح ارزیابی پیشنهادی، I_P تنها به حضور یا فقدان داده های تاریخی (نکات مهم) در مجموعه داده های آموزشی برای یک گزینه و کاربر از مجموعه داده های تست بستگی دارد. با درنظرگیری اینکه نتیجه شاخه مبتنی بر محتوا همیشه درجه ای است که یک گزینه و کاربر معین بیش از یک نکته مهم دارد، پوشش کلی پیشنهادات به شاخه مبتنی بر همکاری بستگی ندارد.

در شکل ۶، ما پوشش پیشنهادات را به عنوان تابعی از زمان برای نشان دادن این امر رسم کرده ایم که چگونه پوشش به نرمی افزایش می یابد وقتی که تعداد نکات مهم شروع به بزرگتر شدن می کند (شکل ۵a) و نسبت میان گزینه ها/کاربران قدیمی و جدید تثبیت می شود.

از روی این تصویر، ما می توانیم همچنین ببینیم که ضمن اینکه یک مقدار پوشش خوبی درباره محلهای رستورانها وجود دارد (۸۵ درصد در سال ۲۰۱۴)، پوشش درباره کاربران کاملا پایین است (۴۵ درصد در ۲۰۱۴). این پوشش پیشنهادات پایین روی کاربران اساسا به دلیل دو علت می باشد. از یک سو، تعدادی کاربران جدید وجود دارد که طی چارچوب زمانی مجموعه داده های تست از Foursquare خارج شده اند (۲۰۱۳ الی ۲۰۱۴). از سوی دیگر، برخی کاربران دقیقا یک بار درباره رستورانهای منهتن نکات مهمی دارند، ولیکن ممکن است رستورانهایی با نکات مهم خارج از منهتن وجود داشته باشد. این پوشش پیشنهادات ضعیف روی کاربران می تواند با سازش سیستم خزش سازی برای همچنین خزش سازی نکات مهم از کاربران به جای خزش سازی نکات مهم از محل ها کاهش یابد.

۳-۵-صحت عملکرد

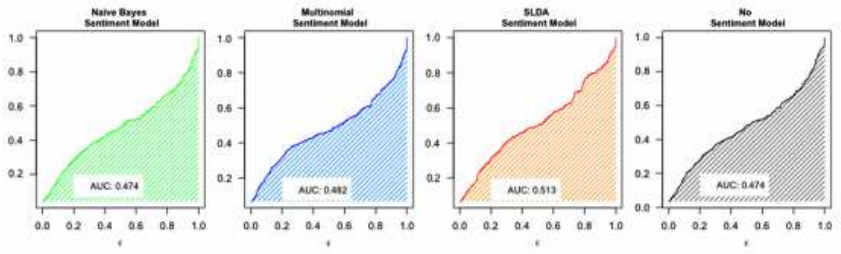
صحت عملکرد خوبی پیشگویی درجه بندی را طبق درجه واقعی اندازه گیری می کند. معمولا، به عنوان تعداد موارد موفقیت آمیز روی همه موارد در مجموعه داده های تست فرموله شده است:

$$accuracy = \frac{\text{good cases}}{\text{overall cases}}. \quad (7)$$

یعنی صحت برابر است با موارد خوب تقسیم بر موارد کل.

جدول ۲-تعداد درجه بندی های پیش بینی شده مثبت و منفی

Rating	SLDA	Multinomial	Naive Bayes	No Sentiment
positive ratings	191.809	142.758	205.585	212.947
negative ratings	21.138	70.189	7.362	0



شکل ۸-AUCها از cdf خطای پیشنهادات برای مدلهای گرایشات مختلف در یک مجموعه داده های متعادلسازی شده

به عنوان تفاوتی با اغلب روشهای ارزیابی در سیستم های پیشنهاددهنده، شیوه ما مفروض می دارد که یک پیشنهاد موفقیت آمیز نوعی است که باعث می شود کاربر به یک تجربه مثبت منجر شود. مثبت بودن تجربه از طریق فیدبک کاربر در مرور اندازه گیری می شود، در حالیکه پیشنهادات با موتور ایجاد می شود.

چون کل تعداد موارد می تواند به موارد مثبت و منفی تجزیه شود، صحت می تواند همچنین از لحاظ متوسط درجه خطای $\overline{\epsilon}$ به ترتیب ذیل بیان شود:

$$accuracy = 1 - \frac{\text{negative cases}}{\text{overall cases}} = 1 - \overline{\epsilon} \tag{8}$$

که در آن ϵ_k میزان خطای پیشنهادات برای آزمایش kام می باشد که طبق تعریف ذیل می باشد:

$$\epsilon_k = \begin{cases} \frac{N_{pos_k}}{N_k} & \text{Sentiment is positive} \\ 1 - \frac{N_{pos_k}}{N_k} & \text{Sentiment is negative} \end{cases} \tag{9}$$

که در آن N_k تعداد گزینه های توصیه شده در k آزمایش و N_{pos_k} موقعیت در یک فهرست مرتب سازی شده است که گزینه ای را اشغال می کند که کاربر تجربه کرده است.

توجه داشته باشید که اگر یک مرور مثبت باشد، آنگاه ϵ_k پایین است اگر گزینه رتبه ای در بالای فهرست پیشنهادات داشته باشد، درحالیکه اگر یک مرور منفی باشد، ϵ_k پایین است اگر گزینه رتبه ای در پایین فهرست داشته باشد.

۱-۳-۵-نتایج مدل مبتنی بر همکاری

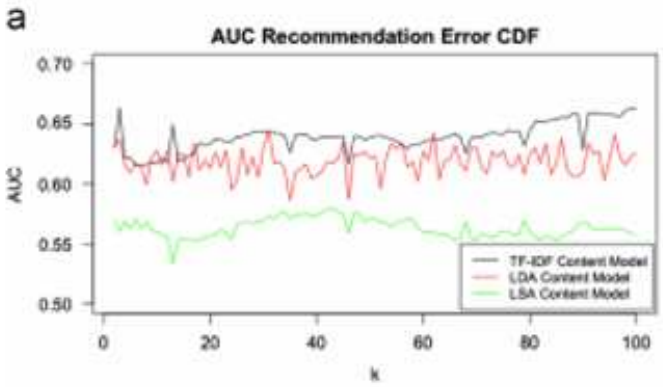
مدلهای مبتنی بر همکاری مطرح شده عقیده جمعیت را با میانگین گیری فیدبک از کاربرانی جمع آوری کرده است که سلایقشان مشابه با کاربر پیشنهاددهنده در محله می باشد. همانگونه که قبلا معرفی گردید، N مجاور از طریق الگوریتم k نزدیکترین مجاور شناسایی شده است که از شباهت کسینوسی تجربیات کاربر تاریخی استفاده کرده اند.

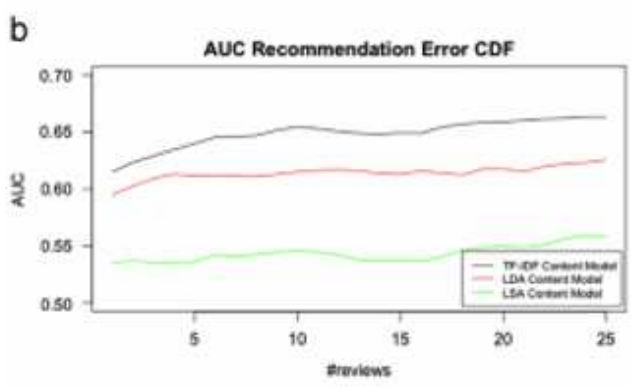
ابتدا، ما مساحت زیر منحنی یا AUC را در توابع ترامک جمعی خطای پیشنهادات cdf رسم می کنیم که به ما اجازه تفسیر خوبی مدلها را با مجموعه های داده های موجود Foursquare می دهد.

طبق شکل ۷ محفظه های ساده از مدلهای چندنامی و SLDA بهتر عمل می کنند. حتی این حقیقت مرتبط تر است که استفاده از هیچ گرایشاتی بهتر از استفاده از یک مدل گرایشات عمل می کند. مدل مبتنی بر همکاری بدون گرایشات همه نکات مهم را به شکل مثبت درنظر گرفته به جای اینکه از یک مدل گرایشات استفاده کند که عقاید را از مرور کننده استخراج می کند.

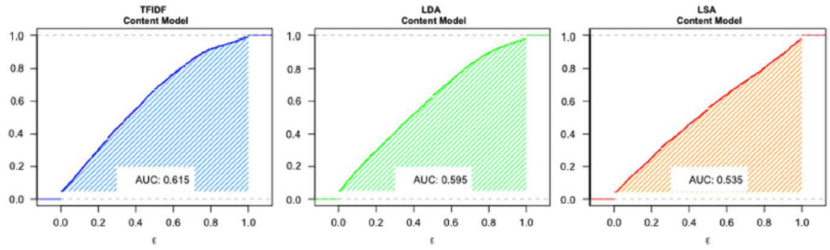
ولیکن، این حقیقت استفاده از مجموعه داده های نامتعادل با تجربیات مثبت بیشتر نسبت به تجربیات منفی به مدلهایی نفع می رساند که به سمت درجه بندی مثبت سوگیری داشته اند. طبق جدول ۲، محفظه های ساده و مشهودا مدل بدون گرایشات دارای نسبت های بزرگتر درجه بندی های مثبت در مجموعه داده های اموزشی می باشد (چرا که طبقه بندی کننده گرایشات به این شیوه ایجاد شده است) و انها در مجموعه داده های تست بهتر عمل می کنند که تعداد تجربیات مثبت بیشتری دارد.

برای کم کردن اثرات عدم تعادل که قبلا توضیح داده شد، ما نمونه گیری فرعی مجموعه داده های تست را به دو مجموعه داده متعادل سازی شده مطرح کرده ایم و بعد صحت عملکرد را محاسبه کرده ایم.





شکل ۹-عملکرد یک شاخه مبتنی بر محتوا. a) AUCی cdf به عنوان تابعی از تعداد ویژگی ها و b) AUC ی cdf به عنوان تابعی از تعداد مرورها



شکل ۱۰-AUC از cdf ی خطای پیشنهادات

برای تخمین مناسب تابع تراکم جمعی و امار AUC، ما فرایند نمونه گیری فرعی را صدها بار تکرار کردیم و نتایج را میانگین گیری کردیم. نتایج این ارزیابی برای هر یک از مدلهای گرایشات در شکل ۸ نشان داده شده است. نتایج این تحلیل نشان می دهد که مدلهای گرایشات بهتر از مدل مبتنی بر همکاری بدون گرایشات عمل کرده است بجز برای محفظه های ساده که مشابها رفتار کرده است. بعلاوه، ما درک کرده ایم که مدلهای SLDA که کلمه چندمعنایی را توجیه می کنند بهتر از سایر موارد نیز عمل کرده اند که در پیشنهاددهنده مبتنی بر همکاری استفاده شده است.

۲-۳-۵-نتایج مدل مبتنی بر محتوا

مدلهای مبتنی بر محتوا با سلایق کاربر سازگاری دارد که توسط ویژگی های استخراجی از داده های نکات نمایش داده می شود. ما پیشتر سه مدل پیشرفته متنی را برای نمایش سلایق کاربر به ویژگی های داده ها ارائه کرده ایم. توجه داشته باشید که مدل محتوایی یک دانش قبلی را درباره یک رستوران جمع اوری می کند. بعبارت دیگر، شباهت میان یک کاربر و یک محل را براساس ترجیحات آنها و صفات آنها به ترتیب توجیه می کند. در نتیجه، ما

مدل محتوایی را طبق این حقیقت ارزیابی کرده ایم که کاربر به محل پیشنهاد شده رفته یا نرفته، که بیش از ارزیابی تجربه در آنجا می باشد، چرا که درک می کنیم که هدف مدل مبتنی بر محتوا تطابق مشخصات مشابه بیش از توجیه عقیده جهانی درباره یک رستوران می باشد.

در آنچه در پی می آید، ما صحت عملکرد این سه مدل را با بررسی AUC یcdf خطای پیشنهادات مقایسه کرده ایم. در این بخش، خطای پیشنهادات به صورتی تعریف می شود که انگار همه نکات در مجموعه داده های تست تجربیات مثبتی هستند.

شکل ۹a نشان دهنده صحت عملکرد به عنوان تابعی از اندازه فضای ویژگی می باشد. توجه داشته باشید که در TF-IDF ویژگی ها کلیدواژه ها، در LSA، مفاهیم معنایی، و در LDA موضوعات اسناد هستند. هرچند پیچیدگی که متضمن مدلهای LSA و LDA می باشد، ساده سازی مدل محتوای TF-IDF تمایل به ایجاد پیشنهادات صحیح تری برای اندازه فضای ویژگی از ۲ الی ۱۰۰ ویژگی داشته است چرا که از یک AUC بزرگتر بهتر عمل کرده و با اینحساب خطای پیشنهادات کمتری دارد.

علی رغم سروصدای این منحنی ها، که به وضوح به مرورهای متنی نویزدار می باشد، یک تمایلی وجود دارد که هرچه تعداد ویژگی ها بیشتر باشد، صحت عملکرد بهتر و AUC بیشتری وجود دارد. ولیکن، اندازه فضای ویژگی مستقیما بر زمان محاسبه اثر دارد و از اینرو بر سرعت انجام پیشنهادات هم اثر دارد. از حالا به بعد، ما از k=۱۰۰ برای همه سه مدل استفاده می کنیم که تضمین کننده یک تبادل جفتی بین صحت عملکرد و هزینه محاسبه می باشد. cdf متفاوت و AUC آنها برای هر یک از مدلهای محتوایی با تعداد k=۱۰۰ ویژگی می تواند در شکل ۱۰ دیده شود.

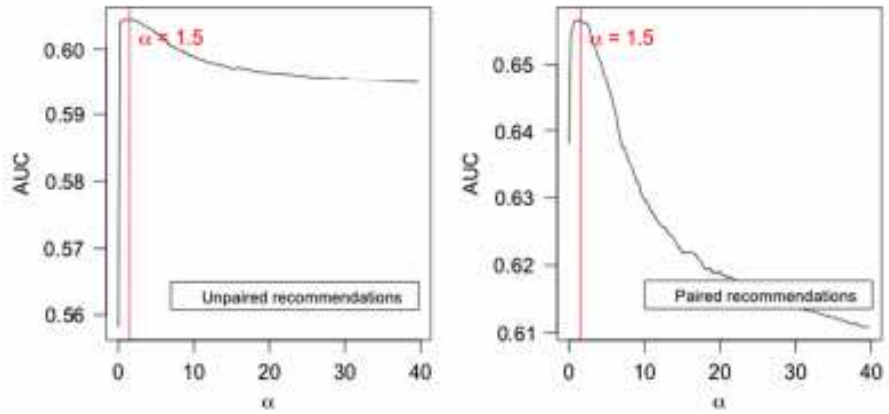
یکی از جنبه هایی که می تواند بر صحت عملکرد مدلهای مبتنی بر محتوا اثر بگذارد تعداد مرورها یا نکات مهم برای هر محل و کاربر است. طبق شکل ۹b، صحت عملکرد با تعداد مرورها به ازای هر محل و هر کاربر افزایش می یابد. نمودارهای گرافیکی AUC برای مشاهدات کاربران و محل ها در مجموعه داده های تست با بیش از یک تعداد معین از مرورها می باشد. در نتیجه، می گوید که هرچه دانش بیشتری درباره محل و کاربر داشته باشیم، بهترمی توانیم پیشنهاداتش را تخمین بزنیم. توجه کنید که TF-IDF از LSA و LDA برای هر تعداد نکات مهم بهتر عمل می کند.

در کل، می توان تفسیر کرد که نه تنها تعداد مرورها بر صحت عملکرد و پوشش یک سیستم پیشنهادات مبتنی بر مرور اثر می گذارد، بلکه سایر ویژگی ها درباره داده هایی نظیر کیفیت مرورها، ناهمگنی موضوعات مرور شده، تاثیر مرورگر در میان سایر ابعاد تاثیر دارد.

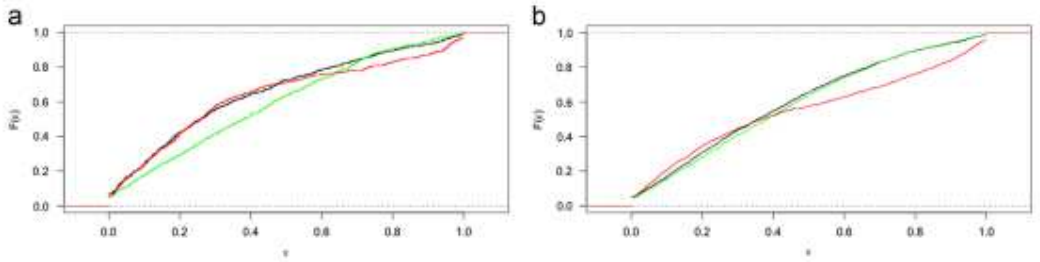
به عنوان اختتام این بخش اشاره می کنیم که باور داریم انتخاب ما درباره سنجش ها برای کسب کیفیت رتبه بندی های ما برای کار و داده های در دست کافی می باشد. خواننده باید توجه داشته باشد که ما درجه بندی ها را پیشگویی نمی کنیم و از اینرو سنجشهای متداولتری مانند MAE یا RMSE در زمینه کاری ما مناسب نیستند. به رفرانس ۲ و ۲۷ برای بحث عمیق بیشتر در زمینه موضوع روشهای سنجش صحت طبقه بندی برای سیستم های رتبه بندی مراجعه شود.

۳-۳-۵-GeoSRS : سیستم پیشنهاد دهنده هیبرید ما

سرانجام اینکه، ما نتایج GeoSRS را نشان داده ایم که با ترکیب خطی هر دو شیوه مبتنی بر محتوا و مبتنی بر همکاری بدست آمده است و طبق بحث ۴,۱ می باشد. هدف ما در این بخش دو جنبه دارد، اول اینکه ما می خواهیم ضریب همبستگی الفا را از معادله ۶ بهینه سازی کنیم که خطای پیشنهادات را به حداقل خودش می رساند. به عبارت دیگر، ما در جستجوی یافتن مقادیر الفا می باشیم که AUC ی تابع تراکم جمعی خطای پیشنهادات را به حداکثر برساند. با اینحساب، ما چندین توصیه را برای مقادیر الفای مختلف شبیه سازی کرده ایم و AUC مربوطه اش را محاسبه کرده ایم تا یکی که انرا به حداکثر می رساند بیابیم. دوم اینکه، ما هدفمان نشان دادن این امر است که شیوه هیبرید همواره بهتر از هر مولفه منفردی عمل می کند.



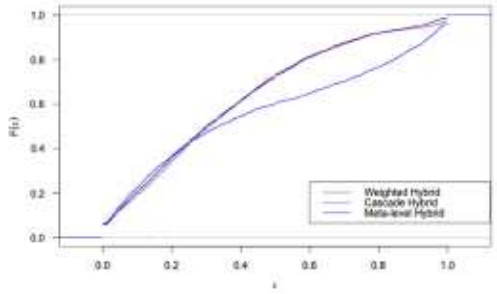
شکل ۱۱-بهینه سازی ترکیب کننده خطی هیبرید



شکل ۱۲-توابع تراکم جمعی تجربی برای تنظیمات مبتنی بر محتوا (سبز)، مبتنی بر همکاری (قرمز) و هیبرید (سیاه). a) پیشنهادات جفتی و b) پیشنهادات غیرجفتی. (برای تفسیرات مرجع ها به رنگ در این زیرنوس شکل، خواننده را به نسخه وب این مقاله راهنمایی می کنیم).

جدول ۳-AUC و پوشش برای تنظیمات مختلف پیشنهاددهنده های مبتنی بر مرو

پیشنهادات	هیبرید	مبتنی بر محتوا	مبتنی بر همکاری	درصد پوشش
پیشنهادات جفتی پیشنهادات غیرجفتی	0.6566 0.6044	0.5984 0.5922	0.6351 0.5514	8.34 39.81



شکل ۱۳-مقایسه طبق تکنیک های هیبریداسیون پیشرفته

برای ارزیابی جهانی پیشنهاددهنده، ما مجددا مدل ارزیابی را با تجربیات تست طبقه بندی شده به عنوان مثبت یا منفی درنظر می گیریم. مدل گرایشات SLDA انتخاب می شود چرا که در تست ارزیابی متعادلسازی شده نمره بالاتری دارد. بعلاوه، TF-IDF به عنوان مدل مبتنی بر محتوا انتخاب می شود که در این شیوه هیبرید گنجانده شده است.

شکل ۱۱ رسم دو منحنی است که هر دو نمایانگر AUCی GeoSRS برای مقادیر مختلف α از صفر تا ۲۰ طبق معادله ۶ می باشد. توجه داشته باشید که شاخه مبتنی بر همکاری همیشه از پیشگویی ها به دلیل فقدان مجاوران

نزدیک به کاربر معین بهتر عمل نمی کند. منحنی سمت چپی عملکرد GeoSRS را برای همه بررسی ها در مجموعه داده های تست نشان می دهد درصورتیکه منحنی سمت راستی عملکرد مرورهای GeoSRS را نشان می دهد که می تواند توسط هر دو شاخه های سیستم (پیشنهادات جفتی) تحت پوشش قرار گیرد (یک پیشنهاد انجام دهد). طبق منحنی، یک مقدار α معادل ۱,۵ مقدار AUCها را از هر دو شرح برنامه به حداکثر می رساند. توجه داشته باشید که α=۰ مربوط به پیشنهاددهنده مبتنی بر همکاری می باشد درحالیکه $\alpha \to \infty$ مربوط به پیشنهاد دهنده مبتنی بر محتوا می باشد.

بعد، ما cdf تجربی را برای هردو موقعیت با α=۱,۵ رسم کرده ایم. شکل ۱۲a نشان دهنده صحت عملکرد برای پیشنهادات جفتی میب اشد. این حقیقت که سیستم هیبرید دارای AUC بزرگتری است همچنین به وضوح در جدول ۳ دیده شده است.

شکل ۱۲b تابع تراکم جمعی تجربی را برای مجموعه داده های Foursquare واقعی تحت نفوذ پیشنهادات جفت نشده رسم کرده است که این بدان معناست که توصیه ها الزاما یک پیشگویی مبتنی بر همکاری را برای هر پیشگویی مبتنی بر محتوا ندارند. بویژه، ما مشاهده کرده ایم که در موقعیت ما برای هر ۵ پیشنهاد مبتنی بر محتوا، تنها یک پیشنهاد مبتنی بر همکاری وجود دارد. این امر اثر عظیمی بر ارزیابی سیستم کل دارد چون پیشنهادات جفتی نشده بیش از پیشنهادات جفتی وزن دارد و شاخه مبتنی بر محتوا به نظر می رسد که به طور صحیح تری عمل می کند. همانند شکل ۱۲b، منحنی هیبرید اکنون به منحنی مبتنی بر محتوا نسبت به منحنی مبتنی بر همکاری نزدیکتر است که خیلی پایین تر بوده است چرا که نمی تواند از یک پیشنهاد مناسب برای آندسته کاربران بدون مجاورت بهتر عمل کند.

با تماس نتایج در معرض تقسیم بندی به دو مجموعه داده مجزا (پیشنهادات جفتی و غیرجفتی)، حقیقت استفاده از یک مدل خطی همیشه عملکرد بزرگتر را تضمین می کند. عملکرد پیشنهاد دهنده هیبرید از زمانی سود می برد که شاخه مبتنی بر همکاری می تواند از یک درجه بندی بهتر باشد یا زمانی که نمی تواند. مع ذلک، بزرگترین نقش ها زمانی رخ می دهد که شاخه مبتنی بر همکاری می تواند توصیه کند.

بجا است که بگوییم محدودسازی پیشنهاد دهنده صرفا به آندسته کاربران با پیشگویی همکارانه (پیشنهادات جفتی) یک اثر عظیم بر پوشش کلی پیشنهادات دارد چرا که شاخه مبتنی بر همکاری پوشش پایینی به دلیل نادر بودن

بالای محلهای درجه بندی عمومی دارد. جدول ۳ رابطه میان صحت عملکرد و را از لحاظAUC برای پیکربندی های سیستم مختلف و پوشش پیشنهادات در هر موقعیت مجموعه داده های ذکر شده قبلی نشان می دهد.

بخاطر مقایسه، ما دو تکنیک هیبریداسیون پیشرفته را اجرا کرده ایم که نشان داده شده بویژه هنگام بکارگیری دو مولفه استقامت های مختلف (برای مثال مبتنی بر همکاری و محتوا به نام سطح کلان و ابشاری) خوب کار می کند.

در رابطه با تنظیمات سطح کلان، ما یک شیوه مبتنی بر همکاری را از طریق شیوه محتوایی مشابه با Pazzani می سازیم که از ماتریس توصیه مبتنی بر محتوا برای شناسایی مجاورت نزدیک هر کاربر تحت ارزیابی استفاده می کند. میزان نادیده بوسیله میانگین گیری میزان گرایشات مجاوران شان و با محاسبه قبلی از طریق مدل تحلیل گرایشات تخمین زده می شود. برعکس، پیکربندی ابشاری یک هیبرید سلسله مراتبی مستحکمی را می سازد که ابتدا شاخه قوی تر را بکار می گیرد و در شرح برنامه ها این شکل همکارانه است و بعد از شاخه مبتنی بر محتوا استفاده می کند برای همه گزینه هایی که روش همکاری نمی تواند به خوبی تصمیم بگیرد. یک سیستم خیلی مشابه توسط Burke در رفرانس ۸ برای پیشنهادات رستوران مطرح گردید ولیکن سیستم هیبرید آنها به روش مبتنی بر دانش در محل شاخه پیشنهاد دهنده مبتنی بر محتوا متکی بوده است.

شکل ۱۳ نشان دهنده منحنی های تراکم تجربی خطا برای هیبریدهای توزین شده سطح کلان و ابشاری می باشد. همانگونه که نشان داده شده است، سیستم هیبرید توزین شده ساده استفاده شده در GeoSRS از دو سیستم دیگر بهتر عمل کرده که باعث فعالسازی یک تکنیک هیبریداسیون ساده ولیکن قدرتمندتر می شود.

۶-نتیجه گیری ها و کارآتی

GeoSRS باعث می شود که کاربران شبکه های اجتماعی مسئله بار بیش از حد اطلاعات را با داده کاوی مرورهای متنی و گزینه های شخصی سازی شده پیشنهاد دهنده طبق ترجیحات آنها براساس مرورهای گذشته شان کم سازند. GeoSRS یک تلاش تازه را برای ملحق کردن صرف داده های مرور متنی به سیستم پیشنهاددهنده در حال کار قبول کرده است.

این مقاله به ارزیابی تنظیمات مختلف GeoSRS می پردازد که از تکنیک های متن کاوی پیشرفته استفاده می کند. طبق شیوه ارزیابی افلاین مطرح شده، نتایج با ترکیب محتوای مرور متنی و گرایشات با یک موتور

پیشنهاددهنده هیبرید توجیه می شود. این امر بویژه به مجموعه داده های متعادل و جفتی مربوط است. به طور منسجم، نتایج ما نشان می دهد که برای مجموعه داده هایمان بهترین پیکربندی توسط TF-IDF برای شاخه مبتنی بر محتوا و توسط SLDA برای شاخه مبتنی بر همکاری با یک ضریب همبستگی α=۱,۵ معین می شود. بعلاوه، نتایج ارزیابی به اهمیت فرایند بازیابی داده ها و کیفیت مجموعه داده ها با نشان دادن این امر اشاره دارد که هر چه تعداد مرورها بیشتر باشد، صحت عملکرد هم بیشترا ست. دست کم ما همچنین نشان داده ایم که مزیت های استفاده از تکینک ساده هیبریداسیون مانند ترکیب خطی توزین شده در مقایسه با بکارگیری تکنیک های پیچیده تر مانند آبشاری یا سطح کلان کدام است.

درون این طرح ارزیابی و GeoSRS، ما نشان داده ایم که خوبی پوشش پیشنهادات به این حقیقت داشتن مرورهای تاریخی برای همه کاربران و همه گزینه ها مربوط می شود. ما مشاهده کرده ایم که رشد فعالیت شبکه اجتماعی طی سالیان مستقیما این ارقام ارزیابی را بهبود داده است، ولیکن ما همچنین بیان می کنیم که هرچه داده های بازیابی شده بیشتر باشد، پوشش بهتر است. ما فرضیه می دهیم که کیفیت پیشنهادات ما باید حین اینکه تعداد مرورها طی زمان افزایش می یابد، بهبود یابد.

برای اثر مستقیم روی صحت عملکرد و پوشش پیشنهادات GeoSRS، این مقاله همچنین استفاده از الگوریتم Quadtree را برای بازیابی موثر داده های موقعیت جغرافیایی مانند مرورها از Foursquare بازیابی کرده است. یک نسخه موازی Quadtree در این مقاله با هدف خزش داده های موثر از نواحی بزرگ شهری به طور تفصیلی آورده شده اند. در نتیجه، GeoSRS همراه با فرایند خزش سازی Quadtree مولفه کلیدی یک سیستم پیشنهاد دهنده قادر به کار در مجموعه داده های بزرگ بررسی ها برای نواحی شهری بزرگ می شود.

سرانجام اینکه، ما می توانیم نتیجه گیری نماییم که نتایج GeoSRS همانند بسیاری سیستم های پیشنهاددهنده دیگر، به شدت به موجودیت و کیفیت داده ها بستگی دارد. ما نشان داده ایم که با متعادلسازی مجموعه های داده ها یا دادن پیشنهادات جفتی در تنظیمات هیبرید، نتیجه سیستم می تواند متفاوت گردد. به همین دلیل، ما عمیقا تشویق می کنیم که سیستم GeoSRS از تست های ارزیابی انلاین از جمله ازمایشات تصادفی سازی شده یا تست A/B برای تسکین چولگی روی مجموعه داده های تست Foursquare و گرفتن نتایج قابل اتکاتر بگذرد.

ترجمه فا
TarjomeFa.Com

این مقاله، از سری مقالات ترجمه شده رایگان سایت ترجمه فا میباشد که با فرمت PDF در اختیار شما عزیزان قرار گرفته است. در صورت تمایل میتوانید با کلیک بر روی دکمه های زیر از سایر مقالات نیز استفاده نمایید:

✔ لیست مقالات ترجمه شده

✔ لیست مقالات ترجمه شده رایگان

✔ لیست جدیدترین مقالات انگلیسی ISI

سایت ترجمه فا ؛ مرجع جدیدترین مقالات ترجمه شده از نشریات معتبر خارجی