دانشگاه شهید بهشتی

دانشکده مهندسی و علوم کامپیوتر

پیشنهاد برچسب برای اسناد متنی با به کارگیری معیارهای تازگی و تنوع

پایان‌نامه کارشناسی ارشد مهندسی کامپیوتر
گرایش هوش مصنوعی

نگارش

احمد دری

استاد راهنما

دکتر احمدعلی آبین

تابستان ۱۳۹۸

فهرست مطالب

فهرست شکل‌ها

فهرست جداول

چکیده

به کلیدواژه‌هایی که به صورت دلخواه توسط کاربران وب و نرم‌افزارهای کاربردی تولید و به اشیاء(سند، تصویر، ویدئو و ...) تخصیص داده می‌شوند، برچسب می‌گویند. این کاربران انگیزه‌های گوناگونی از برچسب‌گذاری می‌توانند داشته باشند. اما موضوع مشترک و مورد توجه ما این است که این برچسب‌ها ابرداده ارزشمندی برای توصیف محتواهای موجود در محیط مبدأ خود محسوب می‌شوند. یکی از راهکارهای بالا بردن کیفیت برچسب‌های تولید شده توسط کاربران، ایجاد فضای مناسب برای برچسب‌گذاری است. از جمله کارهایی که برای ایجاد این فضای مناسب می‌تواند انجام شود، پیاده‌سازی یک سیستم پیشنهاد برچسب است. تکنیک‌های پیشنهاد برچسب سعی دارند تا با پیشنهاد لیستی از برچسب‌ها، کاربران را در برچسب‌گذاری راهنمایی کنند. ما این مسأله را در فضای وب سایت Stack Overflow بررسی و مطرح کردیم. Stack Overflow بزرگترین وب سایت پرسش و پاسخ در حوزه برنامه‌نویسی محسوب می‌شود. در این سایت کاربران پرسش‌ها و چالش‌هایشان را در مباحث مختلف برنامه‌نویسی در جهت یافتن راه‌حل‌های مناسب با سایر کابران به اشتراک می‌گذارند. حال فرض کنید می‌خواهیم به کاربران این سایت بعد از نوشتن متن پرسش و پیش از انتشار آن، لیستی از برچسب‌ها پیشنهاد دهیم. روش پیشنهادی در این پژوهش سعی دارد تا گامی در شخصی‌سازی تکنیک‌های پیشنهاد برچسب بردارد، امری که در سیستم‌های توصیه‌گر بسیار مرسوم است ولی در مسأله‌ی پیشنهاد برچسب کمتر به آن پرداخته شده است. روش پیشنهادی شامل مراحل مختلفی است که بخش اصلی آن را یک مدل شبکه عصبی مصنوعی عمیق تشکیل می‌دهد. برای آموزش این شبکه عصبی دو جریان ورودی داده طراحی شده است. متون پرسش‌های کاربران پس از استفاده از تکنیک‌های جاسازی لغات و تبدیل به مجموعه‌ای از بردارها، جریان اول ورودی شبکه را تغذیه می‌کنند. جریان دوم ورودی شامل پروفایل کاربرانی است که آن پرسش‌ها را مطرح و برچسب‌گذاری کرده‌اند. مدل پیشنهادی یاد می‌گیرد که یک کاربر با یک پروفایل (سلیقه) مشخص یک متن جدید را چگونه برچسب‌گذاری خواهد کرد. در نهایت بعد از آموزش مدل، روش‌هایی را ارائه می‌دهیم که در نتیجه آن معیارهای تنوع و تازگی در لیست برچسب‌های پیشنهادی دخیل می‌شوند. با ایجاد دسته‌بندی‌هایی از برچسب‌ها به گونه‌ای که این دسته‌ها متنوع باشند، مدل‌های طراحی شده را جداگانه بر روی هرکدام از این دسته‌بندی‌ها آموزش می‌دهیم. سپس با طراحی یک تابع که پارامترهای تصادفی هم در آن دخیل است از بین خروجی‌های مدل‌های آموزش دیده شده

خروجی نهایی تولید می‌شود. در نهایت نتایج نشان داد که تکنیک مطرح شده با افزایش p@۵ و p@۱۰ نسبت به روش‌های پایه همراه است.

واژگان کلیدی: برچسب، سیستم‌های توصیه‌گر، پیشنهاد برچسب

فصل ۱

مقدمه

با گسترش شبکه‌های اجتماعی و برنامه‌های کاربردی تحت وب شاهد افزایش تعداد کاربرانی هستیم که با وب در تعامل هستند. آن‌ها در شبکه‌های اجتماعی و صفحات مجازی محتوایی اعم از تصویر، متن، ویدئو و غیره به اشتراک می‌گذارند. این افزایش در تعداد کاربران، رشد حجم مطالب تولید شده توسط آن‌ها را هم به همراه دارد و این چالشی را برای سیستم‌های بازیابی اطلاعات ایجاد می‌کند. برای مثال پرس و جو[۱] در بین حجم بالایی از اسناد متنی یا پرس و جو در میان تصاویر موجود در شبکه‌های اجتماعی، نیازمند این است که محتوای ذاتی اسناد یا تصاویر مورد بررسی قرار گیرد تا نتیجه قابل قبولی به کاربر ارائه شود. تحلیل محتوای ذاتی تصاویر، اسناد، ویدئو و غیره کاری زمان بر و دشوار است. یکی از راه‌هایی که سرویس‌های بازیابی اطلاعات از آن بهره می‌گیرند، استفاده از ویژگی‌های متنی‌ای است که همچون یک ابرداده[۲] در کنار محتوای اصلی قرار گرفته‌اند. مانند برچسب‌ها[۳] نظرات[۴]، توضیحات[۵]، عنوان‌ها[۶] و غیره. در این میان برچسب یکی از پر استفاده ترین ویژگی‌ها است [۱].

۱.۱ برچسب و برچسب‌گذاری

به کلیدواژه‌های دلخواهی که توسط کاربران به اشیا(مانند تصویر، سند، ویدئو) تخصیص داده می‌شوند برچسب می‌گویند. در سایتی مانند delicious[۷] کاربران می‌توانند بوک‌مارک‌های [۸] خود را ذخیره کنند و این بوک‌مارک‌ها را توسط برچسب‌ها دسته‌بندی کنند و با دیگران به اشتراک بگذارند. در flickr[۹] کاربران تصاویر را با برچسب ساماندهی می‌کنند. همین طور برچسب‌گذاری مناسب کمک می‌کند تا تصاویرشان توسط مخاطبانی در یک جامعه هدف بیشتر دیده شود. در citeulike[۱۰] کاربران مقالات و مطالب علمی مورد علاقه خود را با استفاده از برچسب‌ها دسته‌بندی می‌کنند [۲]. در LastFM[۱۱] که یک سرویس پخش انلاین موسیقی است، کاربران با

[1] Query
[2] Metadata
[3] Tag
[4] Comment
[5] Description
[6] Title
[7] del.icio.us
[8] bookmark
[9] flickr.com
[10] citeulike.org
[11] last.fm

استفاده از برچسب، علاقه‌مندی‌های خود در گوش دادن به موسیقی را نشان می‌دهند یا موسیقی مورد علاقه خود را با توجه به سبک یا سایر ویژگی‌هایش گروه‌بندی می‌کنند. سایت‌هایی مانند LibraryThing[۱] یا shelfari[۲] به کاربران این امکان را می‌دهند که کتاب‌های موجود در سایت را برچسب‌گذاری کنند. در سایت‌های خبری مانند Digg[۳] و SlashDot[۴] هم این قابلیت وجود دارد که کاربران خبرها و داستان‌های موجود را با برچسب‌های دلخواه ذخیره کنند. همین طور در سایر شبکه‌های اجتماعی مانند Twitter، Instagram و Stack Overflow هم شاهد استفاده گسترده از برچسب‌ها هستیم.

پس تا اینجا دیدیم کاربران می‌توانند انگیزه‌های گوناگونی برای برچسب‌گذاری منابع داشته باشند. گاهی هدف از برچسب‌گذاری این است که کاربر در آینده بتواند از بین اطلاعاتی که ذخیره کرده است سریع‌تر مطلب مورد نظر خود را بیابد. مانند سرویس google keep[۵] که کاربران یادداشت‌های خود را با برچسب‌های دلخواه ذخیره می‌کنند. در مواردی کاربران هنگام اشتراک‌گذاری یک محتوا، برای اینکه به صورت مختصر آن محتوا را برای سایر کاربران توصیف کنند از برچسب استفاده می‌کنند. گاهی کاربران با هدف اینکه مطالب به اشتراک گذاشته شده توسط آن‌ها بیشتر دیده شود، برچسب‌گذاری می‌کنند. در واقع این کار برای جلب توجه سایر کاربران است.

۲.۱ چالش‌ها

این تنوع در کاربرد و استفاده از برچسب‌ها، باعث می‌شود چالش‌های کار در این حوزه متنوع باشد. برای مثال برچسب‌های استفاده شده در youtube دارای شکل‌های گوناگونی هستند. مثلا گروهی از برچسب‌ها سعی در توصیف محتوای ویدئو دارند. گروهی دیگر مانند "like" ، "dislike" یا "fantastic" صرفا احساسات کاربر نسبت به آن ویدئو را بیان می‌کنند. برخی از برچسب‌ها واژه‌های بی‌معنی‌ای هستند که در فرهنگ لغت وجود ندارند و مثلا قراردادی بین کاربر و دنبال‌کننده‌هایش است.

[1] librarything.com
[2] shelfari.com
[3] digg.com
[4] slashdot.org
[5] keep.google.com

حال با توجه به اینکه در چه حوزه‌ای می‌خواهیم سیستم پیشنهاد دهنده برچسب را پیاده‌سازی کنیم، باید تکنیک پیاده‌سازی را با توجه به نیازمندی‌های آن حوزه تنظیم کنیم. مثلا شاید در یک برنامه پیشنهاد برچسب "java" برای متنی که می‌دانیم به طور خاص مربوط به "java9" است بهتر باشد ولی شاید در یک محیط دیگر برعکس این عمل نیازمندی اصلی باشد. همینطور ممکن است در یک موردی برچسبی مثل "like" بی‌اهمیت باشد ولی در جایی دیگر این برچسب مهم باشد. نمونه دیگر از این گونه مثال‌ها می‌توان به برچسب‌هایی مانند "to read" یا "seen live" در محیطی مثل LastFM اشاره کرد.

۳.۱ اهمیت برچسب‌گذاری

تا به اینجا در مورد برچسب، اهمیت آن و کاربردهایش به صورت مختصر صحبت کردیم. همینطور تصویری کلی از چالش‌هایی که می‌تواند در این فضا وجود داشته باشد ارائه کردیم. اکنون می‌خواهیم به این نکته توجه کنیم که هرچه کیفیت برچسب‌گذاری توسط کاربران بهتر باشد سرویس‌های بازیابی اطلاعاتی که از برچسب‌ها به عنوان داده‌های ورودیشان کمک می‌گیرند، نتایج مرتبط‌تری را استخراج می‌کنند. همچنین تجربه کاربری در حین استفاده از سایت‌هایی که برچسب‌گذاری در آن‌ها اجباری است بهبود پیدا می‌کند. علاوه بر سرویس‌های جستجوگر مختلفی که از برچسب‌ها بهره می‌گیرند بعضی از روش‌های طراحی سیستم‌های توصیه گر هم برچسب‌ها را یک ابرداده‌ی مناسب برای کار خود می‌شناسند [۳]. از آنجایی که برچسب‌گذاری برای کاربران بعضاً بی‌مورد قلمداد می‌شود، در نتیجه ممکن است این کار خارج از حوصله کاربر باشد. همچنین اگر کاربری تصمیم به برچسب‌گذاری هم داشته باشد ممکن است برچسب‌های تکراری و هم معنی استفاده کند(برای مثال شاید استفاده از دو برچسب non relational و NOSQL به صورت همزمان جالب نباشد) یا از لغات و عباراتی بهره‌گیرد که بهتر باشد به جای آن‌ها واژه مناسب دیگری انتخاب شود، واژه‌ای که در آن لحظه شاید به ذهن کاربر خطور نکرده باشد و پیشنهاد ما می‌تواند به کاربر یادآور شود که واژه مناسب‌تری برای توصیف محتوایش وجود دارد. پس هدایت کاربر به سمت لیست مناسبی از برچسب‌ها که این لیست تا حد ممکن خلاصه، مفید و متنوع باشد و در برگیرنده محتوای تولید شده توسط کاربر باشد، از دو جنبه اهمیت دارد. از یک سو آن دسته از کاربرانی که قصد برچسب‌گذاری ندارند را ترغیب به استفاده از برچسب می‌کند. این باعث می‌شود از لحاظ کمی حجم مطالب

شکل ۱.۱: سایت Stack Overflow

برچسب‌گذاری شده افزایش بیابد. از سوی دیگر آن دسته از کاربرانی که تصمیم به برچسب‌گذاری دارند را با پیشنهاد واژه مناسب و برچسب‌های کاربردی‌تر در برچسب‌گذاری یاری می‌کند و این باعث می‌شود برچسب‌ها از لحاظ کیفی هم ارتقا پیدا کنند.

یکی از مواردی که پیشنهاد برچسب می‌تواند مفید واقع شود سایت‌های پرسش و پاسخ مانند Stack Overflow[1] است. در این سایت‌ها کاربر پس از اینکه متن پرسش خود را نوشت برچسب‌هایی را انتخاب و آن پرسش را منتشر می‌کند. سپس کاربران دیگر چنانچه پاسخ یا نظری در مورد آن پرسش داشته باشند آن را با سایرین در میان می‌گذارند. حال چنانچه این سایت از سرویس پیشنهاد برچسب استفاده کند پس از اینکه کاربر پرسشی را مطرح کرد و پیش از انتشار آن پرسش، لیستی از برچسب‌های پیشنهادی به کاربر ارائه می‌کند. این برچسب‌ها کیفیت بازیابی حاصل از سرویس‌های جستجویی که بر روی داده‌های این سایت‌ها فعال است را بالا می‌برد. علاوه بر این باعث می‌شود تا پرسش سریعتر پاسخ داده شود. مثلا چنانچه کاربری پرسشی با برچسب‌های "java"، "web" و "spring" منتشر کند با احتمال بالاتری کاربرانی که در زمینه جاوا، وب و به طور خاص در مورد فریم‌ورک Spring مهارت دارند این پرسش را می‌بینند و سریع‌تر به پرسش مطرح شده پاسخ می‌دهند. در نتیجه میانگین زمان پاسخ به سوالات کاهش پیدا می‌کند که این باعث افزایش رضایت کاربران از آن سایت می‌شود.

[1] stackoverflow.com

۴.۱ راهکار پیشنهادی

مراحل شکل گیری ایده‌ی روش پیشنهادی از آنجایی آغاز شد که مشاهده کردیم روش‌های موجود در پیشنهاد برچسب با تکنیک‌های موجود در سیستم‌های توصیه‌گر چندان آشنا نیستند. این ناآشنا بودن به صورت خاص خود را در بحث شخصی‌سازی پیشنهادات بیشتر نشان می‌دهد. منظور از شخصی‌سازی پیشنهادات این است که سعی کنیم برچسب‌هایی را پیشنهاد دهیم که به سلیقه‌ی کاربر هدف در برچسب‌گذاری نزدیک باشد. برای مثال پیشنهاد برچسب جاوا به کاربری که در طی دو سال فعالیتش یک بار هم از این برچسب استفاده نکرده است، احتمالا گزینه‌ي مناسبی نباشد. از این رو همراه با داشتن نگاهی به سیستم‌های توصیه‌گر اقدام به طراحی یک سیستم پیشنهاد برچسب کردیم. در روش پیشنهادی سعی کردیم با به‌کارگیری تکنیک‌های یادگیری ماشین، مدلی را طراحی کنیم که آموزش ببیند برای یک محتوای متنی با توجه به سلیقه کاربر هدف، چه برچسبی را پیشنهاد دهد. مطلب بیان شده توصیفی ساده از راهکار پیشنهادی این پرژوهش بود.

مدل پیشنهادی شامل یک شبکه عصبی مصنوعی است که به عنوان بخش اصلی راهکار پیشنهادی شامل دو جریان ورودی اطلاعات است. جریان اول توسط محتواهای متنی تغذیه می‌شود. این متون پیش از ورود به شبکه با استفاده تکنیک‌های جاسازی لغات[1] به مجموعه‌ای از بردارها تبدیل می‌شوند. جریان دوم توسط پروفایل کاربرانی که محتواهای متنی نام برده را برچسب‌گذاری کرده‌اند، تغذیه می‌شود. این پروفایل در طی فرآیندی بر اساس سوابق کاربر در برچسب‌گذاری ایجاد شده‌است. سیستم نهایی پیشنهاد برچسب ارائه شده در این پژوهش علاوه بر مدل مبتنی بر شبکه عصبی بیان شده شامل پارامترهایی قابل تنظیم است که معیارهای تازگی و تنوع را در پیشنهادات سیستم لحاظ می‌کنند.

۵.۱ ساختار پایان‌نامه

ساختار این پایان‌نامه به این صورت است که در فصل دوم مفاهیم اولیه و مورد نیاز برای درک مطالب مطرح شده در پایان نامه را توضیح می‌دهیم. در فصل سوم به مرور روش‌ها و تکنیک‌های پیشنهاد برچسب می‌پردازیم. فصل چهارم شامل روش پیشنهادی برای ساخت یک سیستم پیشنهاد برچسب است. در فصل پنجم نتایج حاصل

[1] word embedding

از سیستم پیشنهادی را گزارش می‌دهیم. در فصل ششم به جمع بندی مطالب بیان شده می پردازیم و همچنین نکاتی در مورد کارهای آتی پیشنهاد می‌کنیم.

فصل ۲

ادبیات موضوع

در این فصل مفاهیم مورد نیاز در سایر فصول را توضیح می‌دهیم.

۱.۲ سیستم‌های توصیه‌گر

سیستم‌های توصیه‌گر تلاش می‌کنند تا براساس رفتار کاربر سلیقه و علاقه‌مندی کاربر را پیش‌بینی کنند و از میان حجم بالای اشیا موجود در محیط مورد نظر اقلامی را به کاربر پیشنهاد دهد که به نیازمندی او نزدیک باشد. با توجه به حجم بالای محتوای تولید شده در وب وجود چنین سیستم‌هایی به کاربر کمک می‌کند تا سریع‌تر به محتوای مورد نظر خود دست پیدا کند.

۲.۲ کاربر هدف

هنگامی که یک کاربر محتوایی تولید می‌کند و قصد برچسب‌گذاری دارد، در مرحله‌ای که یک سیستم پیشنهاد برچسب قصد پیشنهاد برچسب به آن کاربر را دارد آن کاربر، کاربر هدف خوانده می‌شود.

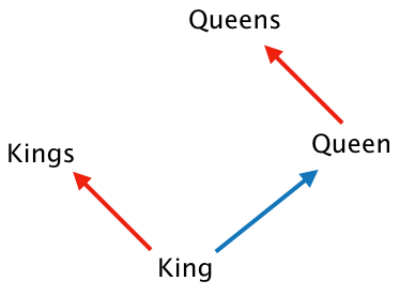
۳.۲ شی هدف

به موجودیتی که توسط کاربر هدف تولید و یا کاربر هدف قصد برچسب‌گذاری آن را دارد، شئ هدف می‌گوییم.

۴.۲ جاسازی لغات

شبکه‌های عصبی یکی از مهمترین ابزار یادگیری ماشین در پردازش زبان طبیعی هستند. یکی از مشکلات مهم استفاده از این شبکه‌ها در کارهای پردازش زبان، نگاشت آن‌ها به بردارهای اعداد است که ورودی شبکه محسوب می‌شوند. به تکنیک نگاشت لغت به بردار در پردازش زبان طبیعی،جاسازی لغت گفته می‌شود. در این روش‌ها، الگوریتم با استفاده از بافتار کلمات سعی بر نگاشتن کلمات به یک فضای جدید را دارد. از بردارهای بدست‌آمده می‌توان به عنوان ورودی‌های شبکه‌ی عصبی استفاده کرد. یکی از ابتدایی ترین روش‌ها برای این‌کار استفاده از بردار یک روشن [1] است. در این روش، اگر n کلمه در پیکره داشته باشیم برای هر کدام از کلمات

[1] one-hot



شکل ۱.۲: کلمه به بردار

یک بردار n بعدی در نظر گرفته می‌شود و در هر بردار فقط یک درایه ۱ می‌شود. ایراد اصلی این روش این است که فضای بسیار زیادی را برای تخصیص بردار هدر می‌دهد. بعلاوه از آنجایی که فقط یک درایه از بردار هر کلمه غیرصفر است، هیچ ارتباط معنایی میان بردار کلمات وجود ندارد.

۱.۴.۲ کلمه به بردار

روش‌های زیادی برای جاسازی لغات ارائه شده است. از آنجایی که در این پایان‌نامه از روش کلمه به بردار [1] استفاده شده است در این قسمت به شرح مختصر آن می‌پردازیم. در سال ۲۰۱۳ یک روش جدید برای جاسازی لغات ارائه شد که در آن کلمات توسط اطلاعات بافتاریشان به فضای برداری نگاشت می‌شوند. در این شبکه‌ی عصبی برای هر کلمه یک بردار تشکیل می‌شود که در واقع نشان دهنده‌ی بافتار آن کلمه است. از آنجایی که کلمات با معنای مشابه در بافتارهای مشابه ظاهر می‌شوند، بردار آن‌ها نیز مشابه می‌شود که نتیجه‌ی آن نگاشت آن‌ها در یک قسمت از فضای بردارهاست. یکی از مزیت‌های اصلی روش جاسازی لغات، این است که از بردارهای بدست آمده می‌توان برای محاسبه‌ی شباهت کلمات استفاده کرد. همانطور که ذکر شد، کلمات مشابه در فضای نگاشت شده در کنار هم قرار می‌گیرند. لذا فاصله‌ی کسینوسی آن‌ها نیز از یکدیگر کم است. بنابراین با استفاده از محاسبه‌ی کسینوسی بردارهای کلمات می‌توان به میزان شباهت معنایی آن‌ها پی برد. شکل ۱.۲ گویای این مطلب است . این شکل نمای ساده‌ی دو بعدی از بردارهای نگاشت شده توسط این روش را نشان می دهد. همانطور که در شکل مشخص است، فاصله‌ی کلمات شاه (king) و ملکه (queen) بسیار شبیه به فاصله‌ی کلمات

[1] word2vec

مرد (man) و زن (woman) است. به عبارت دیگر، این الگوریتم شباهت معنایی بین کلمات مرد و زن را بسیار شبیه به شباهت بین دو کلمه شاه و ملکه تشخیص داده است.

فصل ۳

مروری بر کارهای پیشین