
بسم الله الرحمن الرحیم

دانشگاه یزد

پردیس فنی و مهندسی گروه مهندسی کامپیوتر

پایان‌نامه

برای دریافت درجه کارشناسی ارشد

مهندسی کامپیوتر – نرم‌افزار

عنوان

ارائه توصیه‌گر برخط اخبار مبتنی بر پالایش اشتراکی

استاد راهنما

دکتر سجاد ظریف‌زاده

استاد مشاور

دکتر علی‌محمد زارع بیدکی

پژوهش و نگارش

سیّدعلی الحسینی المدرّسی ی س

اسفند ۱۳۹۵

تحصیلات تکمیلی	**تعهد رعایت حقوق معنوی دانشگاه یزد**	شناسه: ک/۱۳

اینجانب..... دانش‌آموخته مقطع کارشناسی ارشد در رشتهٔ.....
گرایش..... که در تاریخ ..... از پایان‌نامه خود تحت عنوان:.....
با کسب درجه..... دفاع نموده‌ام، شرعاً و قانوناً متعهد می‌شوم:

۱) مطالب مندرج در این پایان‌نامه حاصل تحقیق و پژوهش اینجانب بوده و در مواردی که از دستاوردهای علمی و پژوهشی دیگران اعم از پایان‌نامه، کتاب، مقاله و غیره استفاده نموده‌ام، رعایت کامل امانت را نموده. مطابق مقررات. ارجاع و در فهرست منابع مآخذ اقدام به ذکر آنها نموده‌ام.

۲) تمام یا بخشی از این پایان‌نامه قبلاً برای دریافت هیچ مدرک تحصیلی (هم‌سطح، پایین‌تر یا بالاتر) در سایر دانشگاه‌ها و مؤسسات آموزش عالی ارائه نشده است.

۳) مقالات مستخرج از این پایان‌نامه یا رساله کاملاً حاصل کار اینجانب بوده و از هرگونه جعل داده و یا تغییر اطلاعات پرهیز نموده‌ام.

۴) از ارسال همزمان و یا تکراری مقالات مستخرج از این پایان‌نامه (با بیش از ۳۰ درصد همپوشانی) به نشریات و یا کنگره های گوناگون خودداری نموده و می‌نمایم.

۵) کلیه حقوق مادی و معنوی حاصل از این پایان‌نامه متعلق به دانشگاه یزد بوده و متعهد می‌شوم هرگونه بهره‌مندی و یا نشر دستاوردهای حاصل از این تحقیق اعم از چاپ کتاب، مقاله، ثبت اختراع و غیره (در زمان دانشجویی و یا بعد از فراغت از تحصیل) با کسب اجازه از تیم استادان راهنما و مشاور و حوزه پژوهشی دانشکده باشد.

۶) در صورت اثبات تخلف (در هر زمان) مدرک تحصیلی صادر شده توسط دانشگاه یزد از درجه اعتبار ساقط و اینجانب هیچگونه ادعایی نخواهم داشت.

نام و نام خانوادگی دانشجو:

امضا و تاریخ:

تقدیم به خانواده عزیزم

و تمام کسانی که بدون هیچ چشم داشتی برای سربلندی ایران تلاش می کنند.

از زحمات استاد ارجمندم جناب آقای دکتر ظریف‌زاده که بی شک بدون راهنمایی‌ها و رهنمودهای ایشان تامین این پایان‌نامه میسر نبود، کمال تشکر و قدردانی را دارم.

چکیده

با فراهم شدن دسترسی سریع به گزارش‌های خبری از میلیون‌ها منبع اطلاعاتی در وب، مطالعه‌ی برخط اخبار بسیار رایج شده است. یکی از چالش‌های کلیدی تارنماهای خبری کمک به کاربران در یافتن گزارش‌های جالب و موردعلاقه برای مطالعه است. در این پایان‌نامه، روش جدیدی برای توصیه اخبار پیشنهاد می‌شود که در آن، کاربران بر اساس دو ویژگی برچسب‌های خبری و سرخط خبر به صورت برداری مدل می‌شوند. سپس با استفاده از الگوریتم خوشه‌بندی k-means، کاربران با سلایق خبری مشترک شناسائی می‌شوند. در مرحله بعدی، وزن کاربران نسبت به هم در هر خوشه بر اساس گراف دوبخشی محاسبه و متناسب با این وزن اخبار با بالاترین امتیاز به کاربران توصیه می‌شود. روش معرفی شده با داده‌های واقعی بخش خبری موتور جستجوی پارسی‌جو ارزیابی شده است. نتایج به دست آمده نشان می‌دهد که روش پیشنهادی در مقایسه با الگوریتم‌های دیگر از دقت بالاتری برخوردار است.

کلمات کلیدی: سیستم‌های توصیه، توصیه اخبار، پالایش اشتراکی، شخصی‌سازی.

فهرست مطالب

فهرست جدول‌ها

فهرست شکل‌ها

۱- مقدمه

گسترس انتشار اخبار بر روی اینترنت نیاز به پالایش[۱] آن برای یافتن گزارش‌های جالب و مورد علاقه‌ی کاربران را بیشتر کرده است. تارنماهای تجمیع اخبار[۲]، مانند Google News و Yahoo! News، اخبار را از منابع مختلف جمع آوری نموده و نمایشی کلی از اخبار پیرامون جهان ارائه می‌دهند. مسئله‌ای که چنین سرویس‌های اینترنتی با آن روبه‌رو هستند، حجم بسیار زیاد گزارش‌های خبری[۳] است که کاربران در آن غرق می‌شوند. چالش اصلی کمک به کاربران در یافتن مطالبی است که مطالعه آنها برایشان جالب باشد.

می‌توان گفت که امروزه، کاربران به طور روز افزون با مسئله‌ی سربار اطلاعات[۴] روبه‌رو هستند. سیستم‌های توصیه‌گر[۵] به عنوان وسیله‌ای مناسب جهت برطرف ساختن مسئله‌ی سربار اطلاعات ایجاد شده‌اند. این سیستم‌ها شیء‌های موجود را پالایش می‌کنند که خود منجر به کاهش قابل توجه مسئله‌ی تصمیم‌گیری می‌شود. کاربران به بررسی مجموعه‌ی بزرگ شیء‌ها نمی‌پردازند. در عوض، سیستم‌های توصیه‌گر بخش کوچکی از شیء‌ها را که مرتبط با کاربر می‌پندارند، به آنها ارائه می‌کنند. پژوهش‌ها در زمینه سیستم‌های توصیه‌گر بر روی روش‌های استخراج سلیقه متمرکز بوده است [۱، ۳، ۱۰]. در بحث اخبار، سلیقه‌ی کاربران بیش‌تر به سمت محتوای نهان تا خود شیء‌های خبری گرایش دارد. از این جنبه، توصیه محصولاتی چون فیلم‌ها، آهنگ‌ها یا کتاب‌ها با توصیه گزارش‌های خبری متفاوت است. گزارش‌های خبری شامل نیازمندی‌های خاص خود هستند. این نیازمندی‌ها شامل مجموعه‌ی شیء‌های پویا، مشخصات غیرکامل کاربران و تفاوت‌های بین درگاه‌های خبری مستقل می‌باشد.

مجموعه‌ی شیء‌های پویا به نرخ وارد یا خارج شدن شیء‌ها در سیستم اشاره دارد. سردبیران برای ارائه‌ی آخرین اخبار درباره‌ی رخدادهای اخیر به خوانندگان، شیء‌های خبری تازه را اضافه می‌کنند. از طرف دیگر، گزارش‌های خبری از حیث مرتبط بودن با مرور زمان و آگاهی بیش‌تر کاربران

---

[۱] filter

[۲] News aggreagtion websites

[۳] News article

[۴] Information overload

[۵] Recommender systems

کاهش می‌یابد. مجموعه‌های خبری با نرخ‌های بالاتر افزودن و حذف شیء‌ها در مقایسه با مجموعه-های فیلم و آهنگ روبه‌رو هستند. کاربران ممکن است خواستار استفاده‌ی مجدد از فیلم‌ها و آهنگ-های محبوب خود باشند. در مقابل، خوانندگان به ندرت گزارش‌های خبری قدیمی را مجدداً می-خوانند.

در حالت کلی، با داشتن یک خواننده‌ی اخبار برخط[۶]، اطلاعات پرونده‌ی شخصی[۷] او در ابتدا به وسیله‌ی توصیه‌گر خبر برای توصیف سلایق[۸] مطالب خوانده شده‌ی او جمع‌آوری می‌شود، و بعد از آن گزارش‌های خبری مشخصی از بین نسخه‌های جدید منتشر شده مطبوعات انتخاب می‌شود تا رضایت خواننده را جلب کند. روش‌های سنتی برای پرداختن به مسئله‌ی توصیه اخبار شخصی‌سازی[۹] شده شامل محتوا محور[۱۰]، سیستم‌های پالایش اشتراکی[۱۱] و ویرایش‌های ترکیبی[۱۲] از این دو راه‌حل است. به طور ساده، روش‌های محتوا محور با تطبیق سلایق مطالب خوانده شده‌ی کاربر، امکان پالایش گزارش‌های خبری به سیستم‌های توصیه خبر را می‌دهند. اما توصیه‌گرهای پالایش اشتراکی هدف را تحلیل مطالب خوانده شده‌ی قبلی از کاربران مختلف قرار می‌دهند و بعد از آن گزارش‌های خبری را با استفاده از الگوهای دسترسی مشابه توصیه می‌کنند. همچنین، روش‌های ترکیبی برای کاهش ضعف فردی راه‌حل‌ها مطرح شده‌اند.

علی رغم برخی پیشرفت‌های اخیر، مسئله‌ی توصیه‌ی اخبار شخصی‌سازی شده به سه دلیل کلی همچنان به عنوان چالش باقی مانده است. اولاً، تعداد زیاد مقالات خبری برخط نیازمند مقیاس-پذیری و سرعت عمل بالای سیستم‌های خبری است؛ ثانیاً دریافت دقیق سلایق خواندن از تک تک کاربران امکان‌پذیر نیست چون علاقه‌ی کاربر به مرور زمان تکامل می‌یابد؛ ثالثاً، محبوبیت و تازگی

---

[۶] online
[۷] profile
[۸] preferences
[۹] Personalized news recommendation
[۱۰] Content-based
[۱۱] Collaborative filtering
[۱۲] hybrid

گزارش‌های خبری به طور چشم‌گیری در طول زمان تغییر می‌کند که این امر تفاوت شیء‌های خبری را با محصولات و فیلم‌ها نشان می‌دهد.

عوامل بیان شده باعث می‌شود روش‌های سنتی توصیه، برای مسئله‌ی توصیه‌ی خبر تا حدودی غیرموثر گردد. اگرچه روش توصیه محتوا محور با به کارگیری اطلاعات محتوایی خبرها و اطلاعات گذشته‌ی کاربر سعی می‌کند گزارش‌های خبری را که با سلیقه‌ی کاربر مطابقت حداکثری دارد پیدا کند، اما مدل کردن کاربر تنها بر اساس محتوا برای یافتن علاقه‌مندی‌های وی کافی نیست. برای رفع این مشکل، روش پالایش اشتراکی راه‌حلی است که تلاش دارد تشخیص سلایق کاربر و توصیه‌هایی را که ارائه می‌دهد بر اساس کاربر و داده‌های گروه کاربران به دست آورد. سیستم‌های مبتنی بر پالایش اشتراکی با بهره‌مندی از اطلاعات گذشته‌ی کاربران و اشتراک آنها و همچنین با فرض محتوای تقریباً ثابت، علاقه‌مندی کاربران را به صورت مناسب به دست می‌آورند. بنابراین، این روش در مواردی که محتوای سیستم دائماً در حال تغییر است، موثر نیست.

در این پایان‌نامه، برای مواجه با مشکلات اشاره شده، روش ترکیبی جدیدی مبتنی بر پالایش اشتراکی با استفاده از برچسب[۱۳] اخبار، سرخط[۱۴] اخبار و گراف دوبخشی[۱۵] ارائه می‌شود. در این روش از برچسب اخبار برای بیان محتوای خبر و دریافت سلیقه‌ی کاربران استفاده شده است. استفاده از برچسب اخبار در مقابل محتوای کامل گزارش‌های خبری، کاهش اندازه بردار کاربر را در پی دارد. برچسب‌ها به دلیل ماهیت خود می‌توانند جایگزین بهتری برای بیان علاقه‌مندی‌های کاربران باشد. همچنین با توجه به اینکه محتوای سیستم‌های خبری مدام در حال تغییر است، بیان محتوای خبر به صورت برچسب می‌تواند تا حدودی نرخ تغییرات را تعدیل نماید.

سرخط اخبار، به خبرهای مشابه و حول یک موضوع خبری گفته می‌شود. در روش پیشنهادی، از سرخط اخبار نیز برای دریافت سلیقه‌ی کاربرانی که اخباری مشابه را دنبال می‌کنند،

[۱۳] Tag
[۱۴] Headline
[۱۵] Bipartite graph

استفاده شده است. در نتیجه، کلیک کاربر بر روی یک خبر نشان‌دهنده‌ی علاقه‌مندی وی به برچسب‌ها و سرخط آن خبر نیز در نظر گرفته می‌شود. با استفاده از الگوریتم k-means، کاربرانی که علاقه‌مندی مشابهی در تعقیب برچسب‌ها و سرخط‌ها نشان داده‌اند، در یک خوشه قرار می‌گیرند. سپس در هر خوشه، گراف دوبخشی برای یافتن وزن بین کاربران تشکیل می‌گردد. ارزیابی انجام شده با داده واقعی تأیید می‌کند که الگوریتم پیشنهادی در مقایسه با روش‌های دیگر حداقل دقت را ۳۰ درصد افزایش داده است.

فصل‌های بعدی این پایان‌نامه به شرح زیر می‌باشد: فصل ۲ مروری مختصر بر کارهای قبلی و مرتبط با توصیه‌ی شخصی‌سازی‌شده‌ی اخبار دارد. در فصل ۳ مسئله به صورت رسمی بیان و جزییات الگوریتم توضیح داده می‌شود. نتایج ارزیابی در فصل ۴ گزارش می‌شود و در پایان، فصل ۵ به نتیجه‌گیری می‌پردازد.

۲- مروری بر مطالعات انجام شده

در سال‌های اخیر، مسئله‌ی توصیه در حوزه‌ی اخبار مورد توجه جدی محققان و همینطور شرکت‌های تجاری زیادی قرار گرفته است. از جمله شرکت‌هایی که سرویس توصیه خبری برای تعداد قابل توجهی خواننده‌ی برخط ارائه می‌دهند می‌توان به Google، Yahoo، Swissinfo و DailyMe اشاره کرد. سیستم‌های توصیه‌گر در سه دسته کلی قرار می‌گیرند:

الف) سیستم‌های محتوا محور که از شباهت محتوایی شیء‌های خبری استفاده می‌کند؛

ب) پالایش اشتراکی، که از سلایق مشابه کاربران برای توصیه استفاده می‌کند و

ج) دسته سوم که با ترکیب دو روش قبلی عمل می‌کند [۱۱، ۳-۱].

در این فصل ابتدا علائم استفاده شده در این پایان نامه معرفی، سپس ویژگی‌ها و چالش‌های توصیه‌ی خبرهای شخصی بررسی می‌گردد. در ادامه روش‌های ساده‌ی توصیه ارائه می‌گردد. سرانجام، کارهای انجام شده قبلی به اختصار شرح داده می‌شود.

مفاهیم پایه‌ای که در توصیف الگوریتم استفاده می‌شود در جدول ۲-۱ نشان داده شده است.

جدول ۲-۱ علائم استفاده شده در الگوریتم‌ها

نماد	مفهوم
$\boldsymbol{I}$	مجموعه‌ی شیء‌ها
$\boldsymbol{U}$	مجموعه‌ی کاربران
$\boldsymbol{I(u,i)}$	تابع تعیین تعامل
$\boldsymbol{R}$	ماتریس تعاملات
$\boldsymbol{top(k,c,X)}$	تابع بازگرداندن $\boldsymbol{k}$ تا از بزرگترین مقادیر با توجه به معیار $\boldsymbol{c}$ بر روی مجموعه‌ی $\boldsymbol{X}$

۲-۱- ویژگی‌ها و چالش‌های توصیه اخبار شخصی

در حالت کلی، افراد دسترسی آنی به رخدادهای خبری جدید را ترجیح می‌دهند. انتشارات چاپی قادر به پاسخگویی به این نیاز نیستند. یک راه‌حل خوب مشاهده‌ی گزارش‌های خبری از طریق اینترنت است. در زمان مرور اینترنت، سوالی طبیعی که کاربر با آن مواجه می‌شود چگونگی یافتن رخدادهای خبری جالب توجه در میان انبوه مقالات خبری است [۲، ۵] . برای مواجه با چنین مسئله‌ای، چندین سرویس خبری مشهور مبتنی بر وب، همانند Google News و Yahoo! News، برای ارائه‌ی توصیه خبر برای خوانندگان اخبار برخط بر روی اینترنت ظاهر شده‌اند. معمولاً، سرویس‌های خبری مقالات خبری مرتبط با سلایق خوانندگان را از کاربران مستقل بازیابی می‌کنند، و سرویس‌های خود را با به کارگیری روش‌های توصیه‌دهی مختلف بر اساس تغییر علاقه‌ی مطالب خوانده شده‌ی کاربر تطبیق می‌دهند. این بخش آثار مختلف تعامل کاربران با درگاه-های خبری برخط را معرفی می‌کند. این پدیده‌ها تمایز بین مورد توصیه خبر با دیگر مواردی مانند فیلم‌ها، آهنگ‌ها و کتاب‌ها را معین می‌کند.

سیستم‌های توصیه‌گر موارد استفاده‌ی مختلفی دارند. موارد رایج استفاده شامل تصمیم به مشاهده‌ی کدام فیلم، یا گوش دادن به کدام آهنگ، یا خریدن کدام محصول است. سیستم‌های خبری در این موارد مفید بودن خود را ثابت کرده‌اند [۹]. در مقابل، توصیه خبر با چالش‌های مختلفی روبه‌رو هستند. درباره‌ی پراکندگی، عدم توزان محبوبیت، مجموعه‌های پویای اشیاء، معیارهای محتوایی و دیگر ویژگی‌ها انحصاری گزارش‌های خبری بحث خواهد شد. این جنبه‌ها چالش‌های اصلی برای گرداندگان درگاه‌های خبری که سیستم‌های توصیه‌گر را مدیریت می‌کنند، را نشان می-دهد.

پراکندگی[۱۶]

سیستم‌های توصیه‌گر خبر تعامل کاربران با شیء‌ها را تحت نظر قرار می‌دهند. تعاملات[۱۷] به مجموعه‌ای از حرکت‌ها گفته می‌شود که وابسته به نوع شیء‌ها است. برای مثال، کاربران ممکن است محصولاتی را خریداری کنند، به آهنگی گوش دهند، فیلمی مشاهده کنند یا گزارش خبری بخوانند. این تعاملات را می‌توان با کاردینالیتی[۱۸] مجموعه‌ی کاربران و اشیاء درگیر کمی‌سازی کرد. فرض کنید $u \in U$ و $i \in I$ نشان‌دهنده‌ی کاربران و اشیاء باشد. همچنین، فرض کنید $card(.) = |.|$ نشان‌دهنده‌ی تابعی باشد که تعداد عناصر موجود در مجموعه را برمی‌گرداند. رابطه ۲-۱ پراکندگی را تعریف می‌کند. پراکندگی بیانگر نسبت تعاملات مشاهده شده به کل تعاملات ممکن است. تابع شاخص $I(u,i)$، زمانی که u با i تعامل داشته باشد مقدار ۱ و در غیر این صورت مقدار ۰ برمی‌گرداند.

$$sparsity = 1 - \frac{\sum_{u \in U} \sum_{i \in I} I(u,i)}{|U||I|}$$

رابطه ۲-۱ پراکندگی

$$I(u,i) = \begin{cases} 1 & \text{اگر تعاملی بین u و I صورت گیرد} \\ 0 & \text{درغیراین‌صورت} \end{cases}$$

سیستم‌های توصیه‌گر بر روی محدوده‌های با پراکندگی بالا کار می‌کنند. توصیه کردن شیء‌ها با اطلاعات تقریباً کامل، مسئله‌ای بی‌اهمیت تلقی می‌شود. نبود چینی اطلاعات جامعی، نیاز به مکانیزم توصیه هوشمند را منجر شده است. جدول ۲-۲ سطح پراکندگی مجموعه داده‌های منتخب را نشان می‌دهد. مشاهده می‌شود که اغلب مجموعه داده‌ها شامل کمتر از ۳٪ از تعاملات بالقوه می‌شود. تعاملات بالقوه از ضرب تعداد کاربران و اشیاء به دست می‌آید. علاوه بر این، جدول ۲-۲ رابطه بین

[۱۶] Sparsity
[۱۷] Interaction
[۱۸] cardinality

تعاملات مشاهده شده و تعاملات بالقوه را نشان می‌دهد. برای مثال، مجموعه‌ی داده‌ی Netflix، ۱ در ۸۶,۴ تعاملات بالقوه را نشان می‌دهد. در مقابل، داده‌های ذخیره شده از دو درگاه خبری نشان می‌دهد که این نسبت ۱ در ۶۶۶۲۲,۸ تعاملات بالقوه است. این مهم نشان دهنده‌ی پیچیدگی

مجموعه دادگان	پراکندگی	نسبت تعاملات
Netflix prize challenge	0.98842593	86.4
Book-crossings	0.99998546	68796.6
Movielens 100 k	0.95840128	15.9
Movielens 1M	0.98691797	23.9
Movielens 10M	0.98827612	76.4
EachMovie	0.97631161	42.2
Jester	0.43662440	1.8
Y!Music	0.99915117	1178.8
News Portal 1	0.99998499	66622.8
News Portal 2	0.99996663	2996.8

انتخاب گزارش‌های خبری مناسب به عنوان توصیه می‌باشد.

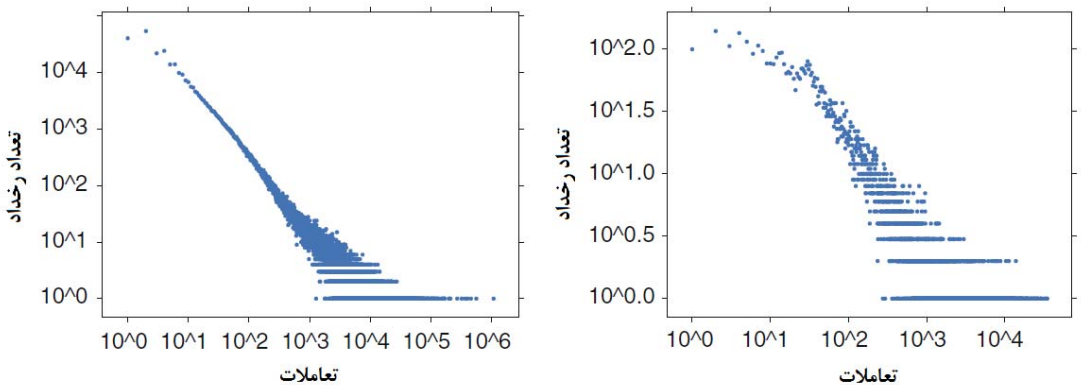
جدول ۲-۲ میزان پراکندگی در متنخبی از مجموعه دادگان معروف [۱]

محبوبیت[۱۹]

از آنجا که بعضی از شیءها با نسبت بالاتری از تعاملات در مقایسه با دیگران روبه‌رو می‌شوند، محبوبیت رخ می‌دهد. کارهای پیشین وقوع عدم توازن محبوبیت در محدوده‌های دیگر را ثبت کرده-اند. این محدوده شامل فیلم، آهنگ و کتاب می‌شود. اسامی متعددی برای محبوبیت شیءها در نظر

[۱۹] Popularity

گرفته شده است. "اقبال"، "موفق"، "پر فروش" به فیلم‌ها، آهنگ‌ها و کتاب‌های محبوب اشاره دارد. سیستم‌های توصیه‌گر این نوع از شیء‌ها را برای توصیه‌ها مناسب می‌دانند. پذیرش توصیه‌ها از شیء‌های محبوب از طرف مراجعه‌کنندگان مورد انتظار است. این پذیرش تا مادامی که سلیقه‌ی کاربر با اغلب کاربران فاصله نداشته باشد، وجود دارد. از طرف دیگر، کاربران ممکن است قبلاً از وجود این شیء‌ها آگاه باشند. در چنین حالت‌هایی توصیه‌گر دچار بدشانسی شده است. عدم توازن محبوبیت با آنالیز توزیع تعاملات در طول زمان بررسی می‌شود. معمولاً محبوبیت، توزیع قانون توان[۲۰] برای تعاملات را نشان می‌دهد. توزیع قانون توان نشان‌دهنده‌ی این است که تعداد کمی از اشیاء نسبت قابل توجهی از تعاملات را در بر می‌گیرند. از طرف دیگر، نسبت زیادی از اشیاء، تعداد نسبتاً کمی از تعاملات را شامل می‌شوند. محبوبیت به عنوان عامل تأثیرگذار در اعتماد کاربران به سیستم توصیه‌گر شناخته شده است. سیستم‌های توصیه‌گری که اشیاء محبوب را توصیه کرده‌اند، شانسی بهتری برای ایجاد تعامل کاربر با سیستم را داشته‌اند. شکل ۲-۱ توزیع محبوبیت یک درگاه خبری همراه با رتبه‌بندی فیلم‌ها در Movielens نشان می‌دهد. این شکل نمودارهای مشابه را برای هر دو، نمایش می‌دهد. تعداد کمی از اشیاء شامل اکثر تعاملات می‌شوند. در مقابل، اکثر شیء‌ها شامل تعداد کمی از تعاملات می‌شوند.



شکل ۲-۱ توزیع محبوبیت برای درگاه خبری (چپ) و MovieLens (راست) [۱]

[۲۰] Power-law

پویای مجموعه‌ی اشیاء

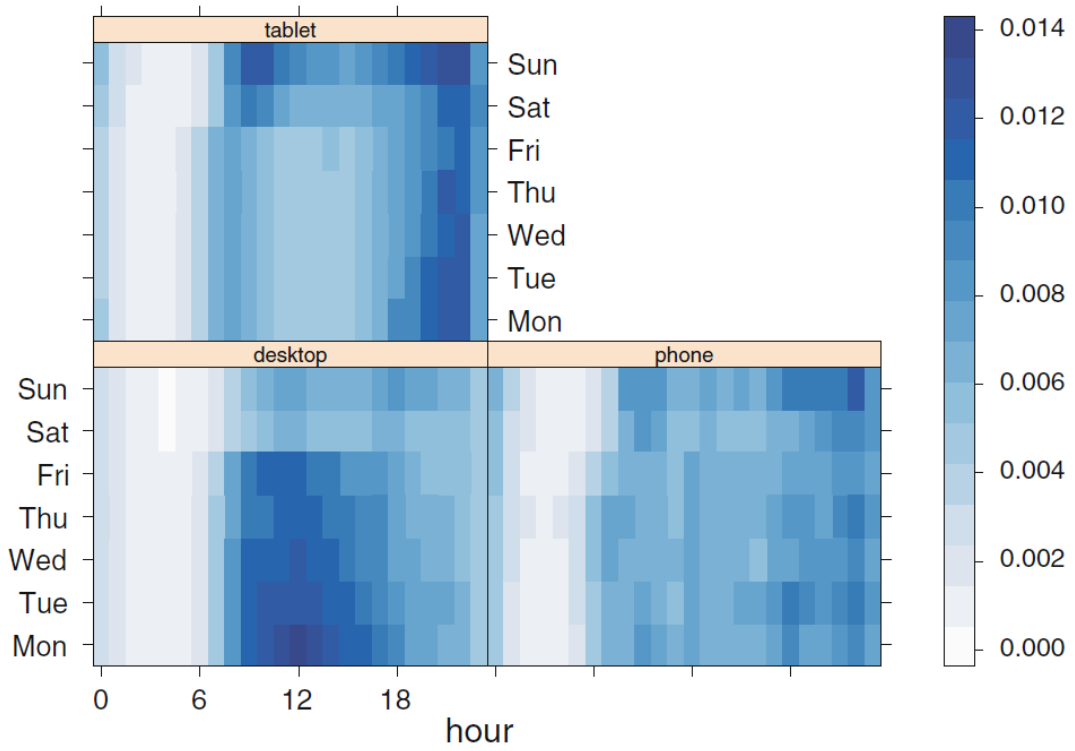
افزودن مدام شیء‌های جدید به مجموعه‌های موجود، یک دلیل اصلی برای سربار اطلاعات را نشان می‌دهد. افزودن‌ها زمانی که یک شرکت فیلم‌سازی فیلمی جدید می‌سازد، ترانه‌ها با آمدن آلبوم جدید یا ناشران کتاب جدیدی را منتشر کنند اتفاق می‌افتد. بعضی از شیء‌های جدید ممکن است محبوب شوند و تعاملات زیادی را جلب کنند. بعضی دیگر به ندرت شناخته می‌شوند. آهنگ ورود شیء‌ها به مجموعه‌ها بستگی به نوع شیء دارد. در مقابل، گزارش‌های خبری معرف شیء‌های با انتشار بالا هستند. درگاه‌های خبری مستقل صدها و هزارها گزارش در سال منتشر می‌کنند.

استفاده و نحوه‌ی مصرف اخبار متفاوت با دیگر حوزه‌ها است. از یک طرف، فیلم‌ها، آهنگ‌ها و کتاب‌ها، کاربران را در دوره‌های زمانی طولانی‌تر جذب می‌کنند. برای مثال، رتبه‌بندی داده‌های مجموعه‌ی داده Movielens را در نظر بگیرید. هر تعامل یک مهرزمانی[۲۱] را همراه دارد. بنابراین، مدت زمان بین آخرین و اولین تعامل برای هر فیلم محاسبه می‌شود. مشاهده می‌شود میانه‌ی مدت زمان‌ها، ۲۲۵۴ روز است. به طور میانگین، گزارش‌های خبری بیش از نیمی از تعاملات خود را در ۲۴ ساعت ابتدایی بعد از انتشار به دست می‌آورند. حتی نسبت تعاملاتی که در ۲۴ ساعت اول متمرکز است، در گزارش‌های خبری محبوب افزایش می‌یابد. این مهم نشان می‌دهد که کاربران از اخبار استفاده‌ی بیش‌تری نسبت به فیلم‌ها دارند. از طرف دیگر، کاربران گاهی کتاب‌ها، فیلم‌ها و آهنگ‌ها را مجدداً استفاده می‌کنند. کاربران با میل خود اقدام به استفاده مجدد می‌کنند. برای مثال، زمانی که به آهنگ محبوب خود گوش می‌دهند یا به تماشای مجدد فیلم محبوب خود می‌نشینند. علاوه بر این، شبکه‌های تلویزیونی آهنگ‌ها و فیلم‌ها را دوباره روی آنتن می‌برند. هیچ شواهدی برای استفاده‌ی مجدد و مکرر کاربران از گزارش‌های خبری وجود ندارد.

[۲۱] Timestamp

معیارهای محتوایی

مصرف اخبار به معیارهای محتوایی مختلفی وابسته است. مصرف اخبار با توجه به زمان روز، روز هفته، مکان و ... متفاوت می‌شود. تعیین محتوای کنونی مسئله‌ای مشکل است. به طور خاص، تعیین اشتباه معیارهای محتوایی مانع از تشخیص صحیح وضعیت‌ها می‌شود. چارچوب محتوایی از ترکیب معیارهای محتوایی تشکیل می‌شود. برای مثال، کاربرانی که اخبار را در طول هفته و بر روی رایانه‌ی شخصی خود می‌خوانند، یک چارچوب محتوایی خاص را نمایندگی می‌کنند. تغییر یک معیار ممکن است منجر به چارچوب محتوایی با نیازمندی متفاوت برای توصیه‌ها شود. برای مثال، کاربرانی که در بعد از ظهر وسط هفته بر روی تبلت در حال خواندن اخبار هستند، ممکن است خواندن گزارش‌های طولانی و جامع را به خاطر محدودیت اندازه صفحه ترجیح ندهند. شکل **۲-۲** به صورت نسبی تعاملات را بر اساس ساعت روز، روز هفته و دستگاه نشان می‌دهد. اکثریت تعاملات ضبط شده برای رایانه‌های رومیزی بر روی ساعت‌های کاری متمرکز است. در مقابل، تلفن‌های همراه و تبلت‌ها نسبت بیش‌تری در بعد از ظهر و آخر هفته به خود اختصاص می‌دهند. به طور کلی، مقدار نسبی کمی از تعاملات در زمان شب برای همه‌ی دستگاه‌ها مشاهده می‌شود. فرض کنید باید الگوریتم توصیه برای درخواستی مشخص انتخاب شود. محتوا جنبه‌ی مهمی است که باید در نظر گرفته شود. درخواست‌ها به احتمال بیش‌تر از طرف دستگاه‌های همراه در آخر هفته داده می‌شود. دستگاه‌های همراه فضای نمایش کمتری برای توصیه فراهم می‌کنند. بنابراین، باید روش توصیه‌ای که در این شرایط بهتر عمل می‌کند، به کار گرفته شود.



شکل ۲-۲ تعداد رخداد نسبی تعاملات در روز، طول هفته و نوع دستگاه [۱]

ویژگی‌های انحصاری گزارش‌های خبری، مانند قالب غیرساخت‌یافته و عمر کوتاه نمایش، موجب تفاوت توصیه خبر با دیگر شیء‌های وب مانند محصولات، فیلم‌ها و افراد می‌شود. در ادامه لیستی از بعضی از مشخصات منحصربه‌فرد شیء‌های خبری بیان می‌شود:

- **حجم انبوه**. متفاوت از دیگر انواع اشیاء وب، گزارش‌های خبری همچون سیلی در دوره‌ی زمانی کوتاه هستند که به محاسبات خیلی بیش‌تری برای توصیه نیاز دارند.
- **قالب غیرساخت‌یافته**. قالب غیرساخت‌یافته یک داستان خبری برای آنالیز نسبت به دیگر اشیاء با ویژگی‌های ساختاریافته مانند محصولات و دوستان مشکل‌تر است.
- **ویژگی‌های موجودیت**. اکثر گزارش‌های خبری وقوع رخداد مشخصی را شرح می‌دهند. خوانندگان اخبار برخط علاقه‌ی بیشتری در اطلاعاتی مانند چه اتفاقی، زمان انجام اتفاق، کجایی آن، کسانی که درگیر بوده‌اند، دارند. به این ویژگی‌ها موجودیت‌های اسمی نیز می-گویند.

- **انتخاب و رتبه‌بندی اخبار**. میزان علاقه به گزاش‌های خبری با توجه به کاربر کاهنده است. به عبارت دیگر، بعد از اینکه فرد روی یک خبری که علاقه دارد، کلیک می‌کند، ارزش علاقه وی ممکن است زمانی که خبر دوم یا بیشتر را کلیک می‌کند کاهش یابد. بنابراین، برای حداکثر کردن رضایت کاربر، رتبه‌بندی اشیاء خبری توصیه شده به توجه ویژه‌ای نیاز دارد.

با این مشخصات ویژه در گزارش‌های خبری، مسئله‌های توصیه خبر شخصی در ذیل خلاصه می‌شود:

مقیاس‌پذیری. مقیاس‌پذیری توصیه خبر به الگوریتم‌های مناسب برای مواجهه بهینه با دادگان انبوه خبری نیاز دارد. راه‌حل‌های مختلفی برای حل مسئله مقیاس‌پذیری استفاده شده است. برای مثال، LSH[۲۲] راه‌حلی بهینه برای جستجوی همسایگی نزدیک در فضای با ابعاد بالا را با روش احتمالی کاهش ابعاد، ارائه می‌دهد [۴، ۷] . Map-Reduce، مدل برنامه‌نویسی است که به وسیله‌ی گوگل پیشنهاد شده است و هدف آن پشتیبانی از محاسبات توزیع شده بر روی دادگان بزرگ به وسیله‌ی خوشه‌های کامپیوترها است و به صورت گسترده در بسیاری کارهای داده‌کاوی و یادگیری ماشین استفاده شده است.

پرونده‌ی شخصی کاربران. پرونده شخصی با کیفیت بالا برای کاربران نمایانگر بهتری از علائق خواندی کاربر است و برای پالایش گزارش‌های خبری برای کاربر خاص بسیار مفید است. معمولاً، برای به دست آوردن اولویت‌های خواندنی کاربر، آنالیزهای دقیق بر روی گزارش‌های خبری غیرساخت‌یافته برای ساخت پرونده‌ی کاربری لازم است. به طور خاص، خواندگان اخبار برخط، اولویت ویژه‌ای برای خلاصه اخبار قائل‌اند. برای مثال چه اتفاقی افتاد، چه زمانی رخ داد، کجا اتفاق افتاد و چه کسانی درگیر بوده‌اند. ساخت پرونده‌ی کاربری ممکن است با آنالیز عمیق‌تر در چنین اطلاعاتی تسهیل شود.

[۲۲] Locality Sensitive Hashing

انتخاب و رتبه‌بندی اخبار. مدل‌های نظری روی توصیه اخبار می‌توانند راه‌حل‌هایی برای مشکل چگونگی انتخاب اشیاء خبری که کاربر به آن علاقه داشته باشد، ارائه کنند. به طور کلی، مدل‌های توصیه بر اساس هم گزارش‌های خبری و هم پرونده‌ی کاربر تولید می‌شود، به طوری که شخصی‌سازی حداکثر شود. همچنین، ویژگی‌هایی نظیر ترتیب لیست اخبار توصیه شده، تنوع اشیاء خبری انتخاب شده به عنوان ملاک‌های مهم دیگری برای یک توصیه‌کننده قوی خبر هستند.

نمایش نتایج. ارائه‌ی مناسب گزارشهای خبری توصیه شده می‌تواند باعث لذت بیش‌تر خوانندگان اخبار شود [۸]. گزارش‌های خبری توصیه شده معمولاً به صورت لیست رتبه‌بندی شده به همراه قطعه‌ی کوچک استخراج شده که محتوای خبر را توصیف می‌کند، هستند. چگونگی افزودن تنوع به اشیاء خبری انتخاب شده به گونه‌ای که افراد بتوانند اطلاعات بیش‌تری با موضوعات متفاوت به دست آورند، یکی از زمینه‌های مناسب برای تحقیق است. به علاوه، ایده‌ی اصلی یک گزارش خبری ممکن است از طریق قطعه‌های کوچک استخراج شده به طور دقیق بیان نشود. راه‌حل دیگری برای نمایش یک شیء خبری شامل استفاده از روش‌های خلاصه‌سازی سند می‌شود تا خلاصه‌ای مختصر و مفید ساخته شود. همچنین، برای ایجاد نتایج توصیه جذاب‌تر می‌توان روش‌های تصویری به کار گرفته شود.

۲-۲- روش‌های ساده‌ی توصیه‌گر خبر

محققان از روش‌های ساده برای مقایسه با روش‌های جدید، استفاده می‌کنند. روش‌های ساده بعضی مزایا به همراه خود دارند. معمولاً، روش‌های ساده پیچیدگی پایینی دارند و پیاده‌سازی آن‌ها آسان است. در بیش‌تر مواقع، روش‌های ساده معیارهای مشخصی را هدف قرار می‌دهند. به عبارت دیگر، روش‌های ساده یک ایده‌ی اصلی را دنبال می‌کنند. در ادامه سه روش ساده توصیه‌گر خبر یعنی

۱. توصیه‌گر محبوب‌ترین شیء
۲. توصیه‌گر جدیدترین شیء
۳. توصیه‌گر تصادفی شیء

معرفی می‌گردد.

۲-۲-۱- توصیه‌گر محبوب‌ترین شیء

توصیه‌گر محبوب‌ترین، شیءها را بر اساس محبوبیت آن‌ها توصیه می‌کند. ایده‌ی اصلی که این روش بر مبنای این است که اشیایی که دارای تعاملاتی با اکثریت کاربران هستند برای دیگر کاربران نیز محبوب خواهند بود. این ایده شبیه مقاله‌های اصلی در روزنامه‌ها و شبیه بخش اصلی در درگاه‌های خبری است. مقاله‌های اصلی توجه بیش‌تری نسبت به مقاله‌های که در پایین روزنامه آمده‌اند، جلب می‌کند. الگوریتم ۱ روند ساخت مدلی که بر اساس توصیه‌گر محبوب‌ترین باشد را شرح می‌دهد. الگوریتم برای توصیه محبوب‌ترین خبر، به ماتریس تعاملات، مجموعه‌ی شیءها و تعداد شیءها به عنوان ورودی نیازمند است. این روش به صورت تکرار، محبوبیت هر شیء را ارزیابی

می‌کند. شیءها به ترتیب k محبوب‌ترین آن‌ها وارد لیست توصیه می‌شوند. الگوریتم می‌تواند با محدود کردن تعاملاتی که دریافت می‌کند، بازه‌های زمانی مختلف را در نظر بگیرد.

الگوریتم ۱ توصیه گر محبوب ترین

ورودی: ماتریس تعاملات R مجموعه اشیاء $\mathcal{I}$ ، تعداد اشیاء برای پیشنهاد k
خروجی: لیست k شیء مرتب شده بر اساس محبوبیت

for all $i \in \mathcal{I}$ **do**
  $popularity(i) \leftarrow \sum_{u \in \mathcal{U}} \mathbb{I}(i, u)$
end for
$recommendations \leftarrow top(k, popularity, \mathcal{I})$

۲-۲-۲- توصیه‌گر جدیدترین شیء

توصیه‌گر جدیدترین بر اساس مفهوم تازگی ساخته می‌شود. همان طور که الگوریتم ۲ نشان می‌دهد، توصیه‌گر جدیدترین، شیءها را بر اساس زمان حضورشان در مجموعه‌ها رتبه‌بندی می‌کند. الگوریتم مجموعه‌ی شیءها، زمان ایجاد شدن، زمان فعلی و تعداد شیءها توصیه را به عنوان ورودی می‌گیرد. این روش اشیایی که باید توصیه شوند را با برش لیست شیءها به ترتیب سن از موقعیت k مشخص می‌کند. با ورود شیءهای جدید به مجموعه، شیءها به بالای لیست می‌روند و با رتبه‌های بالا و قبلی لیست جابه‌جا می‌شوند. بنابراین، این روش شیءهای توصیه را به روز نگه می‌دارد.

الگوریتم ۲ توصیه گر جدیدترین

ورودی: مجموعه اشیاء $\mathcal{I}$ ، مهرزمان ساخت $\tau(i)$ ، زمان فعلی T ، تعداد اشیاء برای پیشنهاد k
خروجی: لیست k شیء مرتب شده بر اساس زمان ساخت

for all $i \in \mathcal{I}$ **do**
  $t(i) \leftarrow T - \tau(i)$
end for
$recommendations \leftarrow top(k, -t, \mathcal{I})$

۲-۲-۳- توصیه‌گر تصادفی شیء

یک روش ساده‌ی دیگر در توصیه، روش توصیه شیءها به صورت تصادفی است. علی رغم اینکه انتخاب تصادفی شیءها خطر توصیه شیءها غیرمرتبط را بالا می‌برد، اما از طرف دیگر، امکان دسترسی به اشیایی که نه محبوب و نه جدید هستند و درنتیجه به وسیله‌ی کاربر یافت نمی‌شدند، را فراهم می‌کند. الگوریتم ۳ روند توصیه‌گر تصادفی شیء را شرح می‌دهد. الگوریتم، به صورت تصادفی شیءها را به لیست توصیه‌ها اضافه می‌کند تا لیست به ظرفیت دلخواه برسد. شیءها نباید تکراری باشند.

الگوریتم ۳ توصیه گر تصادفی

ورودی: مجموعه اشیاء $\mathcal{I}$، تعداد اشیاء برای پیشنهاد k
خروجی: لیست k شیء برای پیشنهاد

while $|recommendations| < k$ **do**
  $i \leftarrow rand(\mathcal{I})$
  if $i \notin recommendations$ **then**
    $recommendations \leftarrow recommendations \cup i$
  end if
end while

۲-۳- روش محتوا محور

از روش محتوا محور به شیوه‌های مختلفی برای ارائه‌ی منتخبی از اخبار شخصی‌سازی شده استفاده شده است [۱،۲]. در سیستم‌های توصیه خبری محتوا محور، از خود خبرها برای محاسبه‌ی شباهت‌ها استفاده می‌شود. با داشتن مجموعه‌ای از گزارش‌های خبری که به تازگی منتشر شده‌اند و اطلاعات گذشته‌ی کاربر، سیستم‌های محتوا محور سعی می‌کنند محتوایی را که با گذشته‌ی کاربر مطابقت دارد پیدا کنند. به طور کلی، محتوای خبری معمولاً در فضای برداری (TF-IDF)، یا با استفاده از توزیع موضوعی که از مدل‌های زبانی به دست می‌آید (PLSI) مدل می‌شود [۷]. پیاده‌سازی سیستم-های توصیه محتوا محور ساده است؛ اما در برخی کاربردها مدل کردن کاربر تنها به صورت مجموعه‌ای از کلمات برای یافتن علاقه‌مندی‌های وی کافی نیست.

پالایش مبتنی بر محتوا بر این فرض استوار است که کاربران به تعامل با اشیایی که محتواهای مشابه دارند، ادامه می‌دهند. برای مثال، کاربران با آهنگ‌ها تعامل دارند. سیستم مشاهده می‌کند که کاربری مکرر به خوانده‌ای خاص مراجعه می‌کند. در نتیجه، سیستم شیءهای مرتبط با این خوانده را توصیه می‌کند. الگوریتم ۷، الگوریتم توصیه مبتنی بر محتوا را نشان می‌دهد. سیستم به مجموعه‌ای از شیءها، ویژگی‌ها و پرونده‌ی شخصی کاربر به همراه تابع شباهت نیاز دارد. الگوریتم شباهت بین همه ترکیبات شیءها را محاسبه می‌کند. در نهایت، پرونده‌ی کاربر به ماتریس شباهت نگاشت می‌شود. در نتیجه، برای هر شیء امتیازی به دست می‌آید. سیستم k شیء بالایی که کاربر هنوز ندیده است را توصیه می‌کند. این روش، بیش‌تر تلاش‌ها را به سمت انتخاب معیار شباهت و همچنین تصمیم به استفاده از کدام ویژگی‌ها سوق می‌دهد.

الگوریتم ۷ پالایش محتوا محور

ورودی: مجموعه اشیاء $\mathcal{I}$ ، ماتریس ویژگی شیء F ، پرونده کاربر U ، تابع شباهت $similarity(X, Y)$ ، تعداد اشیاء برای پیشنهاد k

خروجی: اشیاء مشابه

$S \leftarrow \emptyset$
for all $i \in \mathcal{I}$ **do**
  for all $j \in \mathcal{I} \setminus i$ **do**
    $S_{i,j} \leftarrow similarity(F_i, F_j)$
  end for
end for
$recommendations \leftarrow top(k, \langle U, S \rangle, \mathcal{I} \setminus U)$

۲-۴- روش پالایش اشتراکی

سیستم‌های پالایش اشتراکی از رتبه‌دهی کاربران به شیءها برای ارائه سرویس‌های توصیه استفاده می‌کنند، و به طور کلی، وابسته به محتوای شیءها نیستند. به طور خاص برای توصیه‌ی خبرهای شخصی‌سازی شده، هر گزارش خبری به عنوان یک شیء در نظر گرفته می‌شود و خوانندگان خبری به هر گزارش رتبه می‌دهند. در اینجا، رتبه‌ی شیءها معمولاً به صورت باینری در نظر گرفته می‌شود؛ کلیک کاربر بر روی یک خبر به عنوان رتبه "۱" که نشان‌دهنده‌ی علاقه‌مندی کاربر به خبر

است و کلیک انجام نشده به صورت رتبه "۰" لحاظ می‌شود. در عمل، اکثر سیستم‌های پالایش اشتراکی بر اساس نحوه‌ی رتبه‌دهی قبلی کاربران ساخته می‌شوند. این مهم، یا از طریق استفاده از گروهی از کاربران مشابه با کاربر داده شده برای پیش‌بینی رتبه‌ی اخبار، یا مدل کردن رفتار کاربران بر اساس روش‌های مبتنی بر احتمال انجام می‌گیرد [۶، ۵]. سیستم‌های مبتنی بر پالایش اشتراکی قادرند در حالت‌هایی که میزان اطلاعات گذشته‌ی کاربران در حد مطلوب و دارای اشتراک زیاد و همچنین محتوای سیستم تقریباً ثابت است، رفتار کاربران را به صورت مناسب به دست آورند. اما در بسیاری از مسائل تحت وب، محتوای سیستم دائماً تغییر می‌کند و حتی محبوبیت محتوا نیز در طول زمان متغیر است. درضمن، از بسیاری از کاربران برخط رکوردهای زیادی از گذشته‌ی آنها در دسترس نیست. این موارد جزو ناکارآمدی‌های روش مبتنی بر پالایش اشتراکی محسوب می‌شود.

پالایش اشتراکی مفهوم پیوستگی سلایق مشابه را می‌رساند. به عبارت دیگر، اگر دو کاربر از سلیقه‌های مشابه در گذشته برخوردار باشند، فرض پالایش اشتراکی ترجیح آن‌ها به شیء‌های مشابه خواهد بود. در تحقیقات گذشته الگوریتم‌های فراوانی برای پالایش اشتراکی فراهم شده است. در [۳] تمایز بین پالایش اشتراکی مبتنی بر حافظه و مبتنی بر مدل مشخص شده است. پالایش اشتراکی مبتنی بر حافظه از همه‌ی داده‌های موجود برای توصیه استفاده می‌کند. در مقابل، پالایش اشتراکی مبتنی بر مدل، به الگوهای واضح بین تعاملات کلیت می‌بخشد و توصیه‌ها را بر اساس این مدل‌ها ارائه می‌دهد. روش‌های تجزیه ماتریس در بین بهترین روش‌های پالایش اشتراکی مبتنی بر مدل به حساب می‌آیند.

الگوریتم ۴ توصیه مبتنی بر حافظه از نقطه نظر کاربر را نشان می‌دهد. این روش به مجموعه-ای از کاربران و شیء‌ها، تابع شباهت[۲۳]، تعداد همسایه‌هایی که باید در نظر گرفته شود به همراه طول لیست توصیه که تولید می‌شود، نیازمند است. الگوریتم ابتدا در مجموعه‌ی کاربران می‌چرخد تا تعیین کند سلیقه‌ی چه کسانی به کاربر هدف شبیه است. سپس، این روش ارجحیت شیء‌هایی که

[۲۳] Similarity function

کاربر نسبت به آن‌ها آگاهی ندارد را پیش‌بینی می‌کند. الگوریتم، k شیء با بالاترین امتیاز را بر می‌گرداند.

الگوریتم ۴ پالایش اشتراکی k نزدیکترین همسایگی مبتنی بر کاربر

ورودی: مجموعه کاربران $\mathcal{U}$، مجموعه اشیاء $\mathcal{I}$، تابع شباهت $\sigma(.,.)$، تعداد همسایه l، تعداد اشیاء برای پیشنهاد k

خروجی: لیست k شیء برای پیشنهاد

u
$N \leftarrow \emptyset$
$recommendations \leftarrow \emptyset$
for all $v \in \mathcal{U} \setminus u$ **do**
  $s \leftarrow \sigma(u, v)$
  if $s \geq \sigma(u, N_l)$ **then**
    $N \leftarrow N \cup (v, s)$
  end if
end for
for all $i \in \mathcal{I} \setminus \mathcal{I}_u$ **do**
  for all $n \in N$ **do**
    if $\mathbb{I}(n, i) = 1$ **then**
      $\hat{r}_n \leftarrow s_n r(n, i))$
    end if
  end for
  $\hat{r} \leftarrow \sum_{\mathbb{I}(n,i)=1} \hat{r}_n$
  if $\hat{r} > sort(recommendations_k)$ **then**
    add(i)
    if $|recommendations| > k$ **then**
      $remove(recommendations_k + 1)$
    end if
  end if
end for

الگوریتم ۵ توصیه مبتنی بر حافظه از نقطه نظر شیءها را نشان می‌دهد. در مقایسه با الگوریتم ۴، این روش شباهت بین شیءها را با توجه به تعاملاتشان محاسبه می‌کند. این روش در حالت‌هایی مزیت دارد که $I \ll U$ چون از حلقه‌های با هزینه‌ی محاسباتی بالا بر روی فضای بزرگ کاربری پرهیز می‌شود.

الگوریتم ۵ پالایش اشتراکی k نزدیکترین همسایگی مبتنی بر شیء

ورودی: مجموعه کاربران $\mathcal{U}$، مجموعه اشیاء $\mathcal{I}$، تابع شباهت $\sigma(.,.)$، تعداد همسایه l، تعداد اشیاء برای پیشنهاد k

خروجی: لیست k شیء برای پیشنهاد

u
S
$N \leftarrow \emptyset$
$recommendations \leftarrow \emptyset$
for all $i \in \mathcal{I}$ **do**
  for all $j \in \mathcal{I} \setminus i$ **do**
    $S_{i,j} \leftarrow \sigma(i,j)$
  end for
end for
for all $i \in \mathcal{I}_u^c$ **do**
  $\hat{r}_i \leftarrow u \otimes S_i$
  $recommendations \leftarrow top(k, \hat{r}, \mathcal{I}_u^c)$
end for

روش‌های تجزیه ماتریس به عنوان موفق‌ترین روش‌های پالایش اشتراکی به حساب می‌آیند. این الگوریتم‌ها ماتریس تعامل R، با ابعاد M در N را به تقریب رتبه کم‌تر[۲۴] کاهش می‌دهند. نگاشت کاربران و شیء‌ها به فضای کم‌تر به سیستم‌های توصیه‌گر قابلیت محاسبه‌ی شباهت بین آن‌ها را می‌دهد. الگوریتم ۶ تقریب‌های رتبه‌ی کم‌تر را با تغییر روند کمترین مربعات یاد می‌گیرند. در اینجا، دو ماتریس ویژه به صورت تصادفی مقداردهی اولیه می‌شوند. ابعاد این ماتریس‌ها مطابق تعداد کاربران، شیء‌ها و مقادیر ویژه است. سپس، الگوریتم با تکرار تابع هدف را بهینه می‌کند. تابع هدف میزان نزدیکی تعاملات پیش‌بینی شده با تعاملات مشاهده شده را اندازه‌گیری می‌کند. خطای میانگین مجذورات ریشه[۲۵] (RMSE)، انتخاب متداولی برای چنین تابعی است. الگوریتم یک ویژگی ماتریس را ثابت نگه می‌دارد و در این حین گرادیان ماتریس دیگر را مشخص می‌کند. الگوریتم در گام بعدی تکرار، ماتریس‌ها را تعویض می‌کند. به محض اینکه الگوریتم به معیار توقف می‌رسد، فرآیند با تقریب رتبه کم‌تر به اتمام می‌رسد. معیار توقف شامل حد آستانه‌ها و همچنین بیش‌ترین تعداد تکرار می‌شود. حد آستانه‌ها، حداقل بهبود بین تکرارها را تعریف می‌کند. با مشاهده‌ی بهبود اندک نسبت به میزان تعریف شده، فرآیند به اتمام می‌رسد. در مقابل، بیش‌ترین تعداد تکرار بدون

[۲۴] Lower rank approximation
[۲۵] Root mean squared error

توجه به بهبودها خارج می‌شود. هر دو روش مزیت‌هایی دارند. حد آستانه‌ها، همگرایی به کیفیت مطلوب را تضمین می‌کنند. اما، ممکن است به زمان‌های اجرای طولانی منجر شود. در مقابل، بیش-ترین تعداد تکرار، اجرای محدود را تضمین می‌کند. اما، الگوریتم ممکن است فقط راه‌حل‌های کمتر از حد مطلوب را ارائه دهد. توصیه‌ها با نگاشت کاربر و شیء به زیرفضای رتبه کم‌تر به دست می‌آید.

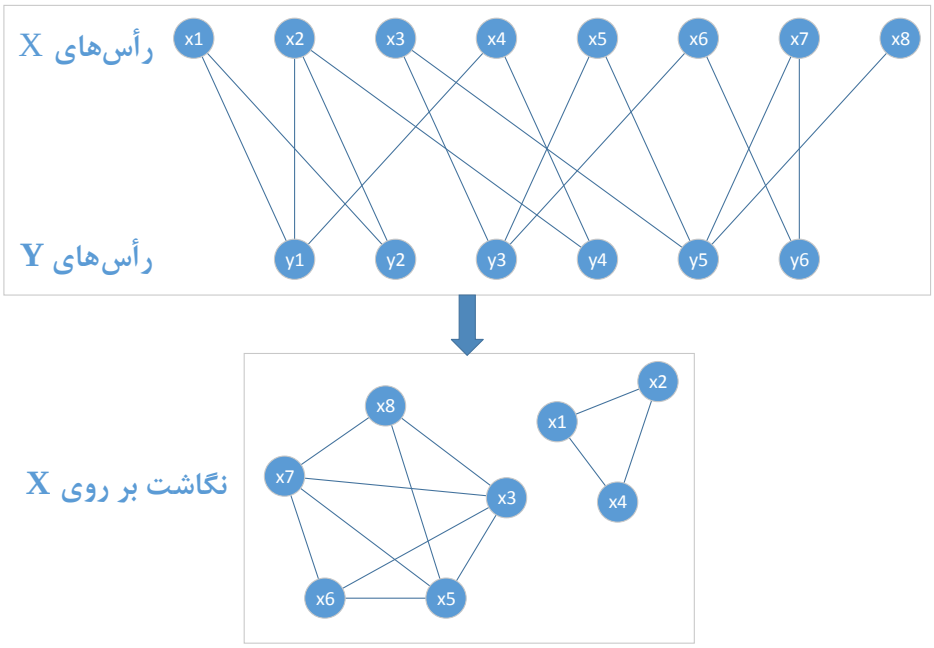
الگوریتم ۶ پالایش اشتراکی تجزیه ماتریس

ورودی: ماتریس تعاملات R ، تعداد مقادیر ویژه در نظر گرفته شده k ، شرط خاتمه ϵ ، تابع بهینه سازی $q(.,.)$
خروجی: تعاملات پیش بینی شده

1: $P \leftarrow rand(|\mathcal{U}|, k)$
2: $Q \leftarrow rand(|\mathcal{I}|, k)$
3: **while** $\epsilon = false$ **do**
4: $\quad P \leftarrow argmax_P q(R, PQ^T)$
5: $\quad Q \leftarrow argmax_Q q(R, PQ^T)$
6: $\quad recommendations \leftarrow top(k, \langle P_u, Q_i \rangle, R)$
7: **end while**

توصیه‌گر مبتنی بر گراف دوبخشی

از جمله روش‌های دیگری که در ذیل پالایش اشتراکی قرار می‌گیرد، روش‌های توصیه بر اساس نظریه گراف می‌باشد. مرجع [۱۲]، روش توصیه‌ی شخصی‌سازی شده مبتنی بر گراف دوبخشی ارائه می‌دهد. گراف دو بخشی، گرافی است که رأس‌های آن به دو مجموعه‌ی X و Y تقسیم می‌شود و ارتباط تنها بین دو رأس در مجموعه‌های متفاوت مجاز است. برای مثال در شکل ۲-۳ می‌توان مجموعه‌ی X و Y را به ترتیب مجموعه‌ی کاربران و اخبار در نظر گرفت. یال‌ها نشان‌دهنده‌ی کلیک کاربران بر روی اخبار است. برای نمایش ارتباط بین رأس‌های مجموعه‌ی X (کاربران)، گراف دوبخشی بر روی مجموعه‌ی X نگاشت می‌شود. نگاشت بر روی مجموعه‌ی X یعنی گرافی که تنها شامل رأس‌های X است و در این گراف دو رأس X به هم متصل است، اگر حداقل یک همسایه مشترک در رأس‌های Y داشته باشند.



شکل ۲-۳ گراف دوبخشی و نگاشت آن بر روی X

بدون از دست دادن کلیت مسئله، هدف یافتن وزن یال‌ها در نگاشت X است. به عبارت دیگر هدف ساختن ماتریس وزن ($w_{n,n}$) است که ارتباط مابین کاربران را بر اساس گراف دوبخشی نشان می‌دهد. در ماتریس وزن، n تعداد کاربران و $w_{i,j}$ اهمیت کاربر i را از منظر کاربر j نشان می‌دهد که به طور کلی با w_{ji} برابر نیست.

نحوه محاسبه‌ی درایه‌های ماتریس وزن در رابطه ۲-۲ آمده است. در این رابطه تابع $k(x)$، درجه رأس x را نشان می‌دهد. a_{il} و a_{jl} درایه‌ی ماتریس مجاورت در گراف دوبخشی هستند. این رابطه نشان می‌دهد، برای حالتی که کاربر j خبرهای کمی را مشاهده کرده باشد و در بین این تعداد خبر اندک اشتراک زیادی با کاربر i داشته باشد، وزن کاربر i از منظر کاربر j بیشتر خواهد بود. و در حالتی که کاربر j خبرهای زیادی را مشاهده کرده باشد و در بین خبرهای مشاهده شده اشتراک زیادی با کاربر i وجود نداشته باشد، وزن کاربر i از منظر کاربر j کمتر خواهد بود.

$$w_{ij} = \frac{1}{k(x_j)} \sum_{l=1}^{m} \frac{a_{il} a_{jl}}{k(y_l)}$$ رابطه ۲-۲

در نهایت، برای توصیه خبر به کاربر i، وزن سایر کاربران از منظر کاربر i به عنوان ضریب امتیاز خبر توصیه شده اعمال می‌شود. در این پایان‌نامه، از الگوریتم گراف دوبخشی در درون خوشه‌ها استفاده شده است که در بخش الگوریتم پیشنهادی توضیح داده می‌شود.

۲-۵- توصیه‌گرهای ترکیبی

همان طور که در بالا توضیح داده شد، سیستم‌های محتوا محور و پالایش اشتراکی قادر به ارائه‌ی توصیه مناسب هستند، اما این روش‌ها دارای نقاط ضعفی نیز هستند. برای به دست آوردن نتایج قابل قبول‌تر محققان امکان ترکیب این دو روش را بررسی نموده، راه‌حل‌های ترکیبی برای توصیه‌ی خبرهای شخصی‌سازی شده را ارائه کرده‌اند.

در این پایان‌نامه یک روش پیشنهادی جدید بر مبنای روش ترکیبی ارائه می‌شود. در روش پیشنهادی از نظرات سایر کاربران و همچنین محتوای داده‌ها برای توصیه‌ی خبر به کاربر استفاده شده است. نوآوری و تفاوت روش پیشنهادی با روش‌های پیشین ارائه شده توسط محققان، در نظر گرفتن برچسب و سرخط اخبار برای در نظر گرفتن محتوای خبر، و همینطور مدل‌سازی و خوشه‌بندی رفتار و سلیقه‌ی کاربران بر اساس برچسب و سرخط خبر می‌باشد.

۳- الگوریتم پیشنهادی توصیه‌گر خبر

مسئله‌ی توصیه‌ی خبرهای شخصی‌سازی شده را می‌توان به این صورت بیان کرد: با داشتن اطلاعات گذشته‌ی N کاربر بر روی M خبر (برای مثال، اطلاعات کلیک کاربران بر روی خبرها)، به طور خاص برای کاربر u، تعداد n خبر خوانده نشده را که دارای بیشترین احتمال علاقه‌مندی از نظر وی باشد توصیه کنید.

برای توسعه و ارزیابی الگوریتم پیشنهادی از مجموعه دادگان بخش خبری موتور جستجوی پارسی‌جو استفاده شده است. این موتور جستجو هم اکنون در مقیاس میلیارد اسناد وب را پوشش و نمایه‌سازی می‌کند. بخش خبری پارسی‌جو، صفحات خبری مربوط به بیش از ۷۰ خبرگزاری اصلی در ایران را خزش کرده و سپس به صورت خودکار، اطلاعات مربوط به بخشهای مختلف خبر را (از جمله عنوان خبر، خلاصه و متن خبر، دسته و برچسب خبر، عکس و فیلم‌های موجود در صفحه) از این صفحات استخراج و پردازش می‌نماید. در این پایان‌نامه، از داده‌های بین ۱۴ ام تا ۱۷ ام شهریور ماه سال ۱۳۹۵ استفاده شده است که این مجموعه شامل تقریبا ۱۵۰۰۰ کاربر منحصربفرد و ۱۱۰۰۰ خبر منحصربفرد می‌باشد (میانگین تعداد کلیک به ازاء هر کاربر نیز در طول یک روز تقریبا ۱۰ بوده است). این مجموعه دادگان، به صورت رکوردهایی به ترتیب حاوی اطلاعات زمان ثبت رکورد، شناسه کاربر، شناسه خبر، شناسه سرخط، دسته‌ی خبری (مثلا سیاسی، ورزشی و غیره) و برچسب‌های خبری است. مدل داده‌ای این مجموع دادگان در جدول جدول ۳-۱ آمده است. شناسه کاربر یک رشته منحصربفرد است که در دفعه اول بازدید کاربر از سایت به وی تعلق گرفته و در کوکی مرورگر وی ذخیره می‌شود. این شناسه هیچگاه منسوخ نمی‌شود، مگر آنکه کاربر آن را پاک کند. سرخط اخبار یا همان مجموعه خبرهای مرتبط با یکدیگر نیز با شباهت‌سنجی متنی خبرها در فیلدهای عنوان، خلاصه و متن خبر به صورت خودکار بدست می‌آید. همانند خود خبرها، هر سرخط خبر نیز دارای یک شناسه منحصربفرد است. در نهایت، دسته‌ی خبر و برچسب‌های آن با ترکیب داده‌های خبرگزاری و روش‌های خودکار تحت نظارت اپراتور انسانی بدست آمده است (صفحه خبر در بسیاری از خبرگزاری‌ها حاوی اطلاعات معتبر مربوط به دسته و برچسب خبر می‌باشد).

جدول ۳-۱ مدل داده‌ای مجموعه دادگان

نام متغیر	نوع متغیر
زمان ثبت رکورد	Datetime (Unix format)
شناسه کاربر	String
شناسه خبر	int
شناسه سرخط	int
دسته‌ی خبر	string {sport, social, politics, international, gallery, economy, culture, health, it, video, science, misc.}
برچسب‌های خبر	Array of strings

با توجه به اطلاعات موجود در مجموعه دادگان و روش پیشنهادی که در این فصل به آن پرداخته می‌شود، هر کاربر به صورت برداری مدل شده است. به طور کلی، به هر کاربر دو نوع بردار شامل بردار سرخط و بردارهای برچسب اختصاص داده شده است. بردار سرخط نشان‌دهنده‌ی سرخط‌های خبری است که کاربر دنبال می‌کند. کلیه‌ی سرخط‌های خبری از روی مجموعه‌ی دادگان مشخص است و مطابق آن در صورتی که کاربری یک سرخط خبری را دنبال (یا کلیک) کند، درایه متناظر با این سرخط در بردار سرخط کاربر از صفر به یک تبدیل می‌شود. بنابراین، همان طور که در رابطه ۳-۱ نشان داده شده است، درایه‌های بردار سرخط کاربر عددهای باینری (صفر و یک) می‌باشند.

$$user_i = \begin{matrix} headline_1 & headline_2 & ... & headline_m \\ [1 & 0 & ... & 1] \end{matrix}$$

رابطه ۳-۱ بردار سرخط کاربر

درایه‌های بردارهای برچسب کاربر نشان‌دهنده وزنی است که هر کاربر به برچسب‌ها اختصاص می‌دهد. هر مقاله‌ی خبری در یکی از ۱۲ دسته موجود (ورزشی، اجتماعی، بین‌المللی، سیاسی، اقتصادی، فرهنگی، فناوری، پزشکی، چندرسانه‌ای و غیره) قرار می‌گیرد. برای هر دسته، برچسب‌های خبری موجود در آن دسته به صورت جداگانه در نظر گرفته می‌شود. بنابراین، زمانی که کاربری بر روی خبری کلیک می‌کند، با توجه به دسته‌ی خبری و برچسب‌های موجود در آن خبر، بردار برچسب کاربر برای آن دسته‌ی خبری به‌روز می‌شود. برای مثال، کاربر بر روی خبر ۱۲۳ کلیک می‌کند. خبر ۱۲۳ در دسته‌ی ورزشی قرار دارد و برچسب‌های این خبر شامل "فوتبال"، "ورزشگاه آزادی"، "پرسپولیس"، "دربی" و "استقلال" می‌شود. بنابراین، در بردار دسته‌ی ورزشی کاربر، وزن برچسب‌های اشاره شده یک واحد افزایش می‌یابد (برعکس بردار سرخط، درایه‌های بردار برچسب باینری نبوده و عدد صحیح بزرگتر یا مساوی با صفر هستند). رابطه ۳-۲ نمونه‌ای از بردار برچسب کاربر را نشان می‌دهد. در نهایت برای هر کاربر به ازای دسته‌های خبری مختلف بردارهایی که نشان‌دهنده وزن برچسب‌ها است، ساخته می‌شود.

$$user_i = \begin{matrix} tag_1 & tag_2 & \dots & tag_m \\ [3 & 2 & \dots & 5] \end{matrix}$$

رابطه ۳-۲ بردار برچسب کاربر

در الگوریتم پیشنهادی توصیه‌گر خبر، برای تولید توصیه‌ها، در گام اول کاربران بر اساس برچسب و سرخط خوشه‌بندی می‌شوند. در گام دوم در هر خوشه، گراف دوبخشی کاربران و خبرهای کلیک شده در آن خوشه ساخته می‌شود. سپس بر اساس روش توضیح داده شده در فصل ۲-۴، ماتریس وزن کاربران در هر خوشه محاسبه می‌شود (شکل ۳-۱). به این ترتیب روش پیشنهادی به ترکیب دو روش مبتنی بر خوشه‌بندی می‌پردازد:

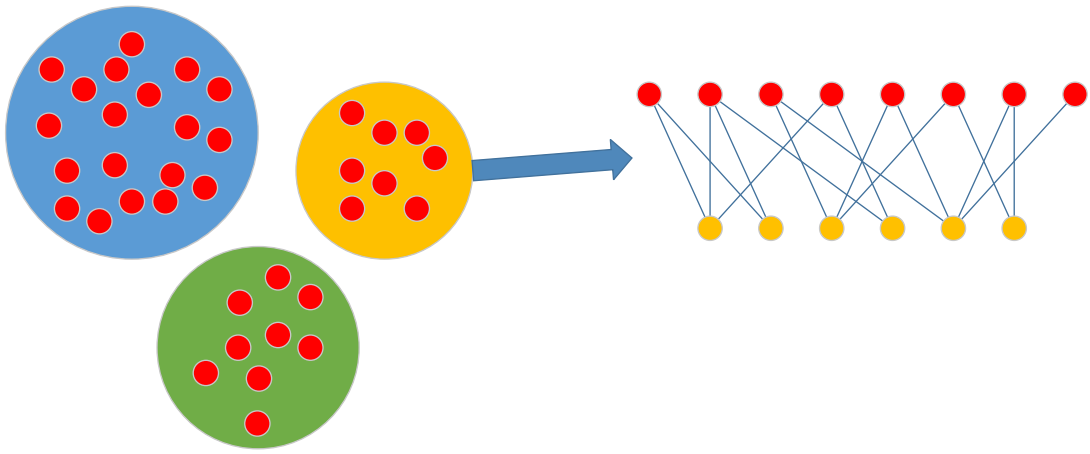
ابتدا کاربران را بر اساس برچسب و سپس آنها را برحسب سرخط خبرها خوشه‌بندی می‌کنیم. در هر یک از این خوشه‌بندی‌ها امتیازی به خبرها تخصیص داده می‌شود. در نهایت، خبرهای با مجموع امتیاز بیش‌تر توصیه‌ی بهتری به حساب می‌آیند. به بیان دقیق‌تر، امتیاز خبر

s_kبرای توصیه به کاربر u_a متناسب با مجموع کلیک‌ها و وزن تمام کاربرانی است که بر روی خبر s_k کلیک کرده‌اند و با کاربر u_a در خوشه یکسان C قرار دارند، یعنی:

رابطه ۳-۳

$$r_{u_a,s_k} \propto \sum_{u_i:u_i \in C} w_{i,a} \times I(u_i, s_k)$$

در فرمول فوق، $I(u_i, s_k)$ برابر با یک است اگر کاربر u_i بر روی خبر s_k کلیک کرده باشد و در غیر این صورت، صفر خواهد بود. $w_{i,a}$ وزن کاربر i از منظر کاربر a را نشان می‌دهد.



شکل ۳-۱ خوشه‌بندی بر اساس کاربران و تشکیل گراف دوبخشی کاربران و اخبار برای هر خوشه

در نهایت، امتیاز داده شده به هر خبر به وسیله‌ی این دو خوشه‌بندی، به صورت میانگین وزن‌دار زیر با هم ترکیب می‌شوند:

رابطه ۳-۴

$$r_{final} = \alpha \times r_{tags} + (1-\alpha) \times r_{headlines}$$

در رابطه ۳-۴، α وزن داده شده به الگورتیم برچسب، r_{tags} امتیاز محاسبه شده درگام یک (خوشه-بندی بردارهای برچسب) و $r_{headlines}$ امتیاز محاسبه شده در گام دوم الگوریتم (خوشه‌بندی بردارهای سرخط) است تا لیست رتبه‌بندی شده از اخبار به دست آید. در نهایت، n تا از خبرهای با بالاترین امتیاز برای توصیه به کاربر انتخاب می‌شوند.

برای خوشه‌بندی کاربران و به دست آوردن خوشه‌ی C که کاربر u_a به آن تعلق دارد، از الگوریتم خوشه‌بندی k-means با معیار فاصله کسینوسی استفاده شده است. رابطه (۳) نحوه محاسبه فاصله کسینوسی بین دو بردار x و y را نشان می‌دهد:

$$cosine\_distance(x,y) = 1 - \frac{x.y}{\|x\|\|y\|}$$ رابطه ۳-۵

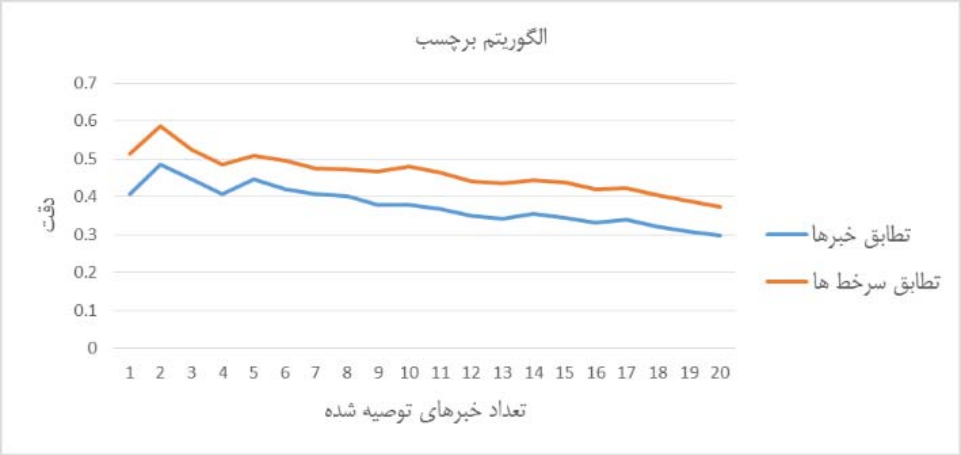
با توجه به رابطه ۳-۵، فاصله کسینوسی برای دو بردار منطبق بر هم، صفر می‌شود. هر چه زاویه بین دو بردار بیش‌تر گردد، این فاصله نیز بیش‌تر می‌شود. در الگوریتم پیشنهادی خوشه‌بندی بر مبنای فاصله کسینوسی مابین کاربران در دو مرحله انجام شده است. برای این منظور در مرحله اول فاصله کسینوسی بردارهای سرخط و در مرحله دوم فاصله کسینوسی بردارها برچسب محاسبه شده است.

۴- تجزیه و تحلیل

برای ارزیابی روش از معیار دقت که در رابطه ۴-۱آمده است، استفاده شده است. مطابق این تعریف، دقت را می‌توان معیار اندازه‌گیری صحیح بودن توصیه در نظر گرفت [۱۰، ۹]. در این رابطه، TP تعداد خبرهای توصیه شده به کاربر است که کاربر روی آن‌ها کلیک کرده و بالعکس، FP تعداد خبرهای توصیه شده به کاربر است که وی روی آن‌ها کلیک نکرده است.

$$precision = \frac{TP}{TP + FP}$$ رابطه ۴-۱

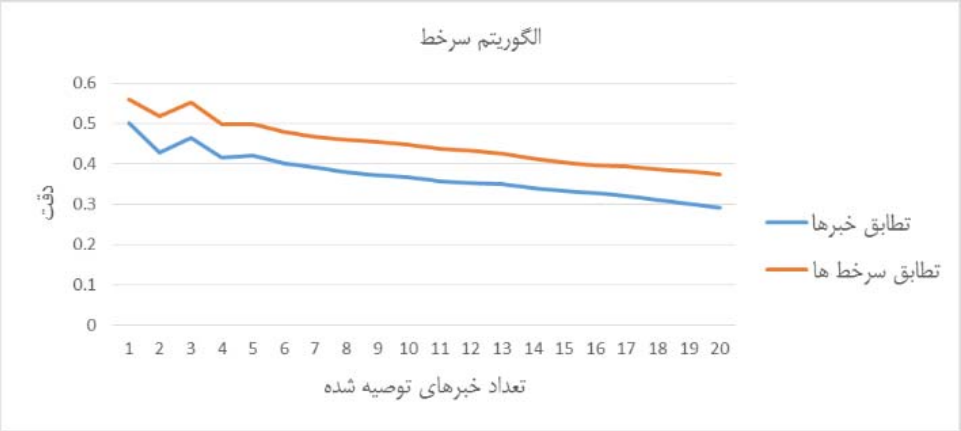
در ارزیابی الگوریتم، داده‌های سه روز اول مجموعه دادگان به عنوان داده یادگیری و داده‌های روز چهارم به عنوان داده آزمایشی در نظر گرفته شده است. بر این اساس، کاربران با داده‌های یادگیری مطابق با روش پیشنهادی خوشه‌بندی می‌شوند و سپس، توصیه اخبار با توجه به خوشه‌های ایجاد شده در روزهای قبل و امتیاز خبرها در روز کنونی انجام می‌گیرد. در نهایت، از داده آزمایشی برای برآورد دقت توصیه استفاده می‌شود. توجه شود که در ارزیابی انجام شده، توصیه اخبار به صورت آفلاین انجام می‌پذیرد (یعنی ابتدا موارد توصیه به کاربر مشخص شده و سپس بررسی می‌شود که کاربر بر روی کدامیک از آنها کلیک نموده است). با توجه به اطلاعات موجود در داده‌های آزمایشی، خبرهای کلیک شده توسط هر کاربر در روز کنونی مشخص بوده و درنتیجه، پارامترهای TP و FP در رابطه ۴-۱ قابل محاسبه خواهند بود.



شکل ۴-۱ دقت الگوریتم برچسب بر حسب تعداد خبرهای توصیه شده

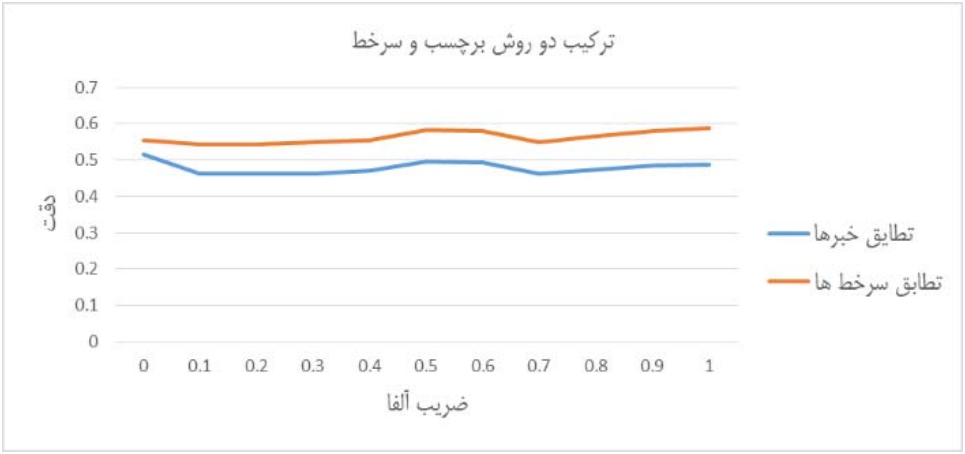
شکل ۴-۱ نتایج دقت را بر حسب تعداد خبرهای توصیه شده در الگوریتم برچسب، یعنی با $\alpha=1$ در رابطه ۳-۴، در قالب دو نمودار تطابق خبرها و تطابق سرخط‌ها نشان می‌دهد. در نمودار تطابق خبرها، خبر توصیه شده عیناً باید توسط کاربر نیز کلیک شده باشد تا به عنوان توصیه صحیح قلمداد شود. ولی در نمودار تطابق سرخط‌ها، کافیست برای یک خبر توصیه شده به کاربر، مابین سرخط آن خبر با سرخط خبرهایی که کاربر کلیک کرده است، تطابق وجود داشته باشد. بدیهی است که در این حالت تعداد خبرهای بیشتری به عنوان توصیه صحیح در نظر گرفته می‌شوند و به همین دلیل، نمودار تطابق سرخط‌ها دقت بیشتری نسبت به نمودار تطابق خبرها از خود نشان می‌دهد. در هر دو نمودار با افزایش تعداد خبرهای توصیه شده دقت کاهش می‌یابد. این اتفاق کاملاً قابل انتظار است، زیرا الگوریتم به صورت آفلاین در حال ارزیابی است و تعداد کلیک‌های هر کاربر در مجموعه داده آزمایشی نیز محدود می‌باشد (برای مثال، حجم قابل توجهی از کاربران دارای تعداد کلیک کمتر از

۴ در روز آزمایش هستند که در این صورت، توصیه بیش از ۴ خبر به آنها نتایج بهتری به همراه ندارد).



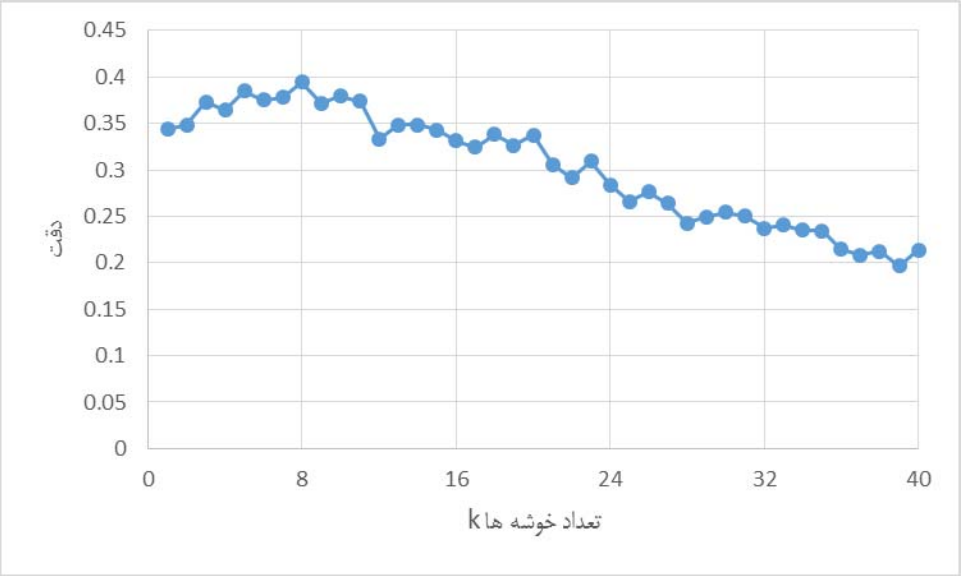
شکل ۴-۲ دقت الگوریتم سرخط برحسب تعداد خبرهای توصیه شده

شکل ۴-۲ نتایج دقت الگوریتم سرخط را (یعنی با $\alpha=0$) بر حسب تعداد خبرهای توصیه شده نشان می‌دهد. هر چند در یک نگاه کلی، این نتایج مشابه با نتایج دقت در شکل ۱ است، ولی الگوریتم سرخط تا حدودی منجر به نتایج بهتری شده است. بنابراین، می‌توان نتیجه گرفت که سرخط خبرها معیار مناسب‌تری برای شباهت‌سنجی کاربران و توصیه اخبار به آنها می‌باشد.



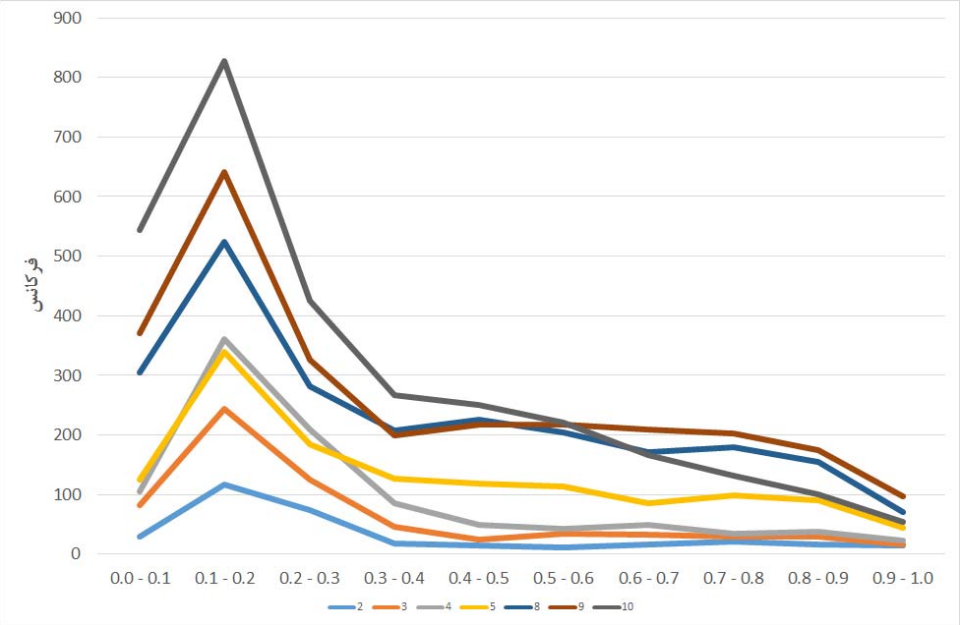
شکل ۴-۳ الگوریتم ترکیبی دو روش برچسب و سرخط

شکل ۴-۳ نمودار ترکیب دو روش برچسب و سرخط را برای مقادیر مختلف α نشان می‌دهد. با در نظر گرفتن دو نمودار تطابق خبرها و تطابق سرخط‌ها، بهترین نتایج با مقدار $\alpha=0.5$ بدست آمده است، یعنی هنگامی که هر دو الگوریتم برچسب و سرخط دارای وزن یکسانی در امتیاز نهایی توصیه خبر در رابطه ۴-۳ بوده‌اند. از همین رو، در ادامه نتایج از همین مقدار برای تنظیم α استفاده می‌کنیم.



شکل ۴-۴ دقت الگوریتم پیشنهادی به ازاء تعداد خوشه‌های مختلف (k)

شکل ۴-۴ نمودار دقت را بر حسب تعداد خوشه‌ها (یا همان پارامتر k) نشان می‌دهد. به صورت کلی می‌توان گفت که به استثنای انتخاب مقادیر کوچک برای k، دقت روش با افزایش پارامتر k روندی کاهشی دارد. طبیعی است که هر چقدر تعداد خوشه‌ها بیشتر شود، تعداد کاربران کمتری در هر خوشه واقع می‌شوند و در نتیجه، طبق پالایش اشتراکی، تعداد خبرهای کمتر و با دقت پایین‌تری برای توصیه وجود خواهند داشت. از طرف دیگر، هر چه تعداد خوشه‌ها بیش از اندازه کاهش یابد، کاربران غیرمشابه بیشتری در یک خوشه جمع می‌شوند که این خود باعث افت دقت توصیه می‌گردد. برای مقایسه روش پیشنهادی با الگوریتم‌های دیگر از $k=8$ که در ارزیابی‌ها منجر به بالاترین دقت شده است، استفاده می‌شود.



شکل ۴-۵ توزیع نسبت فاصله کاربران تا مرکز کلاستر به فاصله تا مرکز کلاسترهای دیگر

برای ارزیابی کیفیت خوشه‌بندی کاربران مطابق با رابطه ۴-۲ برای تمامی کاربران نسبت فاصله‌ی هر کاربر تا مرکز خوشه‌ای که به آن تعلق دارد (d_{i,c_i}) به فاصله تا مرکز کلاسترهای دیگر (d_{i,c_j}) محاسبه شده است.

$$\delta = \frac{d_{i,c_i}}{d_{i,c_j}}$$ رابطه ۴-۲

تعداد رخداد مقادیر محاسبه شده‌ی این نسبت (δ) به فواصل ۰/۱ در شکل ۴-۵ آمده است. توزیع رسم شده نشان می‌دهد خوشه‌بندی کاربران مناسب است زیرا بیشترین فرکانس در بازه‌های کمتر اتفاق افتاده است. یعنی کاربران به درستی به مرکز خوشه خود نسبت به خوشه‌های دیگر نزدیکتر هستند.

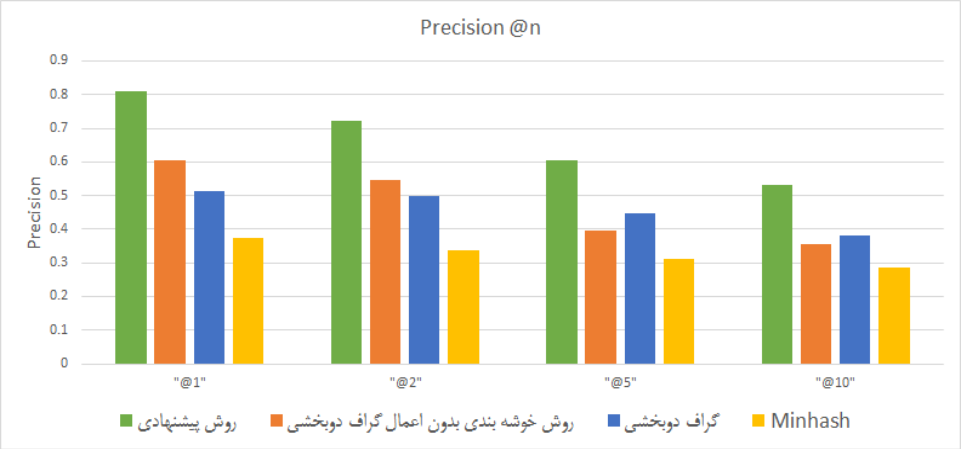
در ادامه، روش پیشنهادی با دو روش توصیه مقایسه و ارزیابی می‌شود. روش اول همان گراف دو بخشی و برگرفته از مقاله [۱۲] است. الگوریتم این روش بر کل کاربران و اخبار اعمال می‌شود. روش دوم بر اساس الگوریتم MinHash و برگرفته از مقاله [۷] است.

الگوریتم MinHash امکان خوشه‌بندی کاربران را بر اساس شباهت ژاکار[۲۶] فراهم می‌سازد. تعریف شباهت ژاکار دو مجموعه A و B (که معادل با مجموعه خبرهای کلیک شده توسط دو کاربر هستند) در رابطه ۴-۳ آمده است:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$ رابطه ۴-۳

برای همه‌ی کاربران، مجموعه خبرهای کلیک شده توسط آن‌ها تشکیل می‌شود و سپس کاربران با استفاده از معیار شباهت فوق خوشه‌بندی می‌شوند. در آخر، خبرهای با بیشترین کلیک در خوشه‌ی کاربر که تاکنون توسط وی مشاهده نشده، برای توصیه به وی انتخاب می‌شوند.

شکل ۴-۶ نمودار مقایسه بین روش پیشنهادی، روش خوشه‌بندی بدون اعمال گراف دوبخشی، روش گراف دوبخشی و الگوریتم MinHash را نشان می‌دهد. در این نمودار، نتایج دقت به ازاء تعداد n خبر توصیه شده به کاربر ترسیم شده است (مقادیر n به ترتیب برابر با یک، دو، پنج و ده در نظر گرفته شده است).



شکل ۴-۶ مقایسه دقت نتایج بین روش پیشنهادی، گراف دوبخشی و MinHash

همانطور که شکل نشان می‌دهد، الگوریتم پیشنهادی بر الگوریتم‌های دیگر برتری دارد. یعنی دقت الگوریتم پیشنهادی بهتر از هر دو روش گراف دوبخشی و MinHash است (یعنی تا ۳۰ درصد بالاتر از گراف دوبخشی و تا ۴۰ درصد بالاتر از MinHash). روش خوشه‌بندی بدون اعمال گراف دوبخشی

[۲۶] Jaccard similarity

حالتی است که در درون خوشه‌ها وزن کاربران بر اساس گراف دوبخشی لحاظ نشود. روش پیشنهادی نسبت به این حالت نیز دقت بالاتری دارد.

۵- نتیجه

در این پایان‌نامه مسئله‌ی توصیه‌ی اخبار شخصی‌سازی شده مورد مطالعه و بررسی قرار گرفت و روش جدیدی برای حل مسئله ارائه شد. در مسئله‌ی توصیه‌ی اخبار برای افراد مختلف باید با داشتن اطلاعات گذشته‌ی N کاربر بر روی M خبر (برای مثال، اطلاعات کلیک کاربران بر روی خبرها)، به طور خاص برای کاربر u، تعداد n خبر را که دارای امکان علاقه‌مندی از نظر وی باشد توصیه نمود.

راه‌حل‌های کلاسیک برای توصیه‌ی اخبار شخصی‌سازی شده به سه دسته عمده تقسیم‌بندی می‌شود:

۱. سیستم‌های محتوا محور که از شباهت محتوایی شیء‌ها خبری استفاده می‌کند؛
۲. پالایش اشتراکی، که از سلایق مشابه کاربران برای توصیه استفاده می‌کند و
۳. دسته سوم که با ترکیب دو روش قبلی عمل می‌کند

دسته‌ی اول یعنی سیستم‌های محتوا محور از نظر فهم و پیاده‌سازی روش‌های آسان و ساده‌ای هستند. در این سیستم‌ها، از خود خبرها برای محاسبه‌ی شباهت‌ها استفاده می‌شود. با داشتن مجموعه‌ای از گزارش‌های خبری که به تازگی منتشر شده‌اند و اطلاعات گذشته‌ی کاربر، سیستم‌های محتوا محور سعی می‌کنند محتوایی را که با گذشته‌ی کاربر مطابقت دارد پیدا کنند. عیب سیستم-های محتوا محور این است که سلیقه کاربران در نظر گرفته نمی‌شود به عبارت دیگر روش‌های بدون بازخورد هستند و از این جهت توصیه‌های آن‌ها از دقت و کارایی لازم برخوردار نیست.

دسته‌ی دوم یعنی سیستم‌های پالایش اشتراکی از رتبه‌دهی کاربران به شیء‌ها برای ارائه سرویس-های توصیه استفاده می‌کنند، و بر خلاف روش‌های دسته‌ی اول توجهی به محتوای خبرها ندارند. برای توصیه‌ی اخبار شخصی‌سازی شده، هر گزارش خبری به عنوان یک شیء در نظر گرفته می‌شود و خوانندگان خبری به هر گزارش رتبه می‌دهند. در اینجا، رتبه‌ی شیء‌ها معمولاً به صورت باینری در نظر گرفته می‌شود؛ کلیک کاربر بر روی یک خبر به عنوان رتبه "۱" که نشان‌دهنده‌ی علاقه‌مندی

کاربر به خبر است و کلیک انجام نشده به صورت رتبه "۰" لحاظ می‌شود. در عمل، اکثر سیستم‌های پالایش اشتراکی بر اساس نحوه‌ی رتبه‌دهی قبلی کاربران ساخته می‌شوند.

تغییرات زیاد سیستم‌های خبری از جمله چالش‌های اصلی در عدم کارایی راه‌حل‌های سنتی مسئله توصیه است. برای آنکه از مزیت توجه همزمان به محتوای خبر و رفتار کاربران بهره برداری شود در این پایان‌نامه یک روش جدبد پیشنهاد شده است. نوآوری و تفاوت روش پیشنهادی با روش‌های موجود، در نظر گرفتن برچسب و سرخط اخبار برای در نظر گرفتن محتوای خبر، و همینطور مدل‌سازی و خوشه‌بندی رفتار و سلیقه‌ی کاربران بر اساس برچسب و سرخط خبر می‌باشد. برچسب اخبار برای بیان محتوای خبر و همچنین تعدیل نرخ تغییرات مناسب می‌باشد و سرخط اخبار هم برای دریافت سلیقه‌ی کاربرانی که روند یکسانی در مطالعه اخبار دارند سودمند است. در روش پیشنهادی، کاربرانی که سوابق مشابهی در تعقیب برچسب‌ها و سرخط‌های خبری داشته‌اند خوشه‌بندی شده‌اند. سپس در هر خوشه گراف دوبخشی کاربران و اخبار تشکیل شده و بر اساس آن وزن بین کاربران محاسبه شده است. در نهایت، امتیاز توصیه اخبار برای هر کاربر بر اساس رفتار کاربران هم‌خوشه با وی و متناسب با وزن محاسبه شده از گراف دوبخشی محاسبه می‌شود. نتایج ارزیابی با یک مجموعه دادگان واقعی برتری دقت روش پیشنهادی را در قیاس با روش‌های گذشته نشان می‌دهد.

بر مبنای بررسی‌ها و پژوهش‌های انجام گرفته در این پایان‌نامه پیشنهادهای ذیل برای ارتقاء و کارایی بیش‌تر سیستم‌های خبرهای شخصی‌سازی شده ارائه می‌گردد:

أ. از روش‌های کاهش اندازه بردار در مقیاس داده بزرگ همچون LSH استفاده شود تا بعد بردار برچسب‌ها و سرخط‌ها کاهش داده شده، سرعت الگوریتم خوشه‌بندی افزایش یابد

ب. الگوریتم پیشنهادی بر روی داده‌های آنلاین اجرا شده و نتیجه‌های جدید با نتایج به دست آمده در این پایان‌نامه مقایسه گردد. این مقایسه می‌تواند در اطمینان خاطر به الگوریتم و توسعه و ارتقاء آن مورد استفاده قرار گیرد.

ج. مدل MapReduce برای برنامه‌نویسی در بخش‌های مختلف نظیر خوشه‌بندی و گراف دوبخشی، به کار گرفته شود. این امر به خصوص برای وقتی که بعد بردارها بسیار زیاد و تعداد خبرها خیلی زیاد می‌شود ضروری است زیرا مدل MapReduce روش مناسبی برای پردازش داده‌های بزرگ می‌باشد.

واژه نامه انگلیسی به فارسی

گراف دوبخشی	Bipartite graph
پالایش اشتراکی	Collaborative filtering
مبتنی بر محتوا	Content-based
پالایش	filter
سرخط	Headline
ترکیبی	hybrid
سربار اطلاعات	Information overload
تعامل	Interaction
شباهت جاکارد	Jaccard similarity
تارنماهای تجمیع اخبار	News aggreagtion websites
گزارش خبری	News article
برخط	online
توصیه شخصی-سازی‌شده‌ی خبر	Personalized news recommendation
محبوبیت	Popularity
قانون توان	Power-law
سلایق	preferences
پرونده‌ی کاربر	profile
سیستم‌های توصیه-گر	Recommender systems
تابع شباهت	Similarity function
پراکندگی	Sparsity
برچسب	Tag
مهرزمان	Timestamp

فهرست منابع و مآخذ

[1] B. Kille, A. Lommatzsch, and T. Brodt, “News Recommendation in Real-Time,” in Smart Information Systems SE - 6, F. Hopfgartner, Ed. Springer International Publishing, 2015, pp. 149–180.

[2] L. Li, D.-D. Wang, S.-Z. Zhu, and T. Li, “Personalized News Recommendation: A Review and an Experimental Investigation,” J. Comput. Sci. Technol., vol. 26, no. 5, pp. 754–766, Sep. 2011.

[3] G. Adomavicius and A. Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 6, pp. 734–749, Jun. 2005.

[4] J. Liu, P. Dolan, and E. R. Pedersen, “Personalized News Recommendation Based on Click Behavior,” in Proceedings of the 15th International Conference on Intelligent User Interfaces, 2010, pp. 31–40.

[5] L. Li, D. Wang, T. Li, D. Knox, and B. Padmanabhan, “SCENE: A Scalable Two-stage Personalized News Recommendation System,” in Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2011, pp. 125–134.

[6] L. Li, W. Chu, J. Langford, and R. E. Schapire, “A Contextual-bandit Approach to Personalized News Article Recommendation,” in Proceedings of the 19th International Conference on World Wide Web, 2010, pp. 661–670.

[7] A. S. Das, M. Datar, A. Garg, and S. Rajaram, “Google News Personalization: Scalable Online Collaborative Filtering,” in Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 271–280.

[8] F. Garcin and B. Faltings, “PEN Recsys: A Personalized News Recommender Systems Framework,” in Proceedings of the 2013 International News Recommender Systems Workshop and Challenge, 2013, pp. 3–9.

[9] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011, pp. 368-369.

[10] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber, “Offline and Online Evaluation of News Recommender Systems at Swissinfo.Ch,” in Proceedings of the 8th ACM Conference on Recommender Systems, 2014, pp. 169–176.

[11] B. Kille, F. Hopfgartner, T. Brodt, and T. Heintz, “The plista dataset,” in Proceedings of the 2013 International News Recommender Systems Workshop and Challenge on - NRS ’13, 2013, pp. 16–23.

[12] T. Zhou, J. Ren, M. Medo, and Y.C. Zhang, “Bipartite network projection and personal recommendation,” Phys. Rev. E, vol. 76, no. 4, p. 46115, Oct. 2007.

Abstract

As the web provides quick access to millions of sources around the world, online news reading has become very popular. A key challenge for news websites is to help users find interesting articles to read. In this paper, we propose a new approach for news recommendation, in which users are modeled as vectors based on news tags and news headlines. Users with similar news preferences are identified by using the k-means clustering algorithm. In the next step, the weight of each user is determined based on the bipartite graph within each cluster and the news with the highest score is recommended to the users. The method is evaluated with actual data from Parsijoo search engine. The results demonstrate that the proposed method has higher accuracy in comparison with other algorithms.

Keywords: Recommender systems, News recommendation, Collaborative filtering, Personalization, Bipartite graph

Yazd University
Faculty Electrical and Computer Engineering

A Dissertation Submitted in Partial Fulfillment of the Requirement
for the Master Degree in Computer Engineering

Title

An Online News Recommender based on Collaborative Filtering

Supervisor
Dr. Sajjad Zarifzadeh

Advisor
Dr. Ali Mohammad Zare Bidoki

By
Seyedali Alhosseini Almodarresi Yasin

March 2017