ISSN: 2745-3673
http://www.Rcsj.ir

مجله‌ی علمی پژوهش در علوم رایانه (سال دوم)
شماره ۴ / زمستان ۱۳۹۵ / ص ۷- ۱

سیستم توصیه‌گر آگاه از متن مبتنی بر گراف

سعیده سادات مومنی[۱]، علی رجائی[۲]
[۱] دانشگاه تربیت مدرس، دانشکده علوم ریاضی
[۲] دانشگاه تربیت مدرس، دانشکده علوم ریاضی

نام و نشانی ایمیل نویسنده مسئول:
سعیده سادات مومنی
Saeedeeh_momeni@yahoo.com

چکیده

حجم اطلاعات در وب و تعداد کاربران اینترنت در سال های اخیر با نرخ بی سابقه ای افزایش یافته است. با این رشد استثنائی تعامل کارآمد کاربر با اینترنت و تشخیص اطلاعات مربوط بسیار مهم می شود. این مشکل که سرریز داده ها نامیده می شود باعث به وجود آمدن سیستم های پیشنهاد دهنده شد. این سیستم ها در میان حجم عظیم اطلاعات با تحلیل رفتار کاربر خود اقدام به پیش بینی و پیشنهاد را در رابطه با مورد مورد نظر کاربر می نمایند]۶[. بنابراین جمع آوری اطلاعات در مورد علایق کاربران بسیار مهم است. بیشتر روش‌های موجود فقط روی دامنه مورد و کاربر تمرکز می‌کنند و هر گونه اطلاعات متنی اضافه مثل زمان و مکان را در نظر نمی‌گیرند. در این مقاله روش‌های مختلف پیاده‌سازی سیستم پیشنهاددهنده‌ی آگاه از متن مورد بررسی قرار گرفته و یک سیستم مبتنی بر گراف برای توصیه موسیقی به کاربران با قابلیت انعطاف پذیری ارائه خواهد شد.

واژگان کلیدی: سیستم پیشنهاد دهنده - آگاه از متن – چند بعدی - گام تصادفی - بازخورد ضمنی

مجلـه‌ی علمی پژوهش در علوم رایانه ، شماره ۴ ،زمستان ۱۳۹۵، ص ۷ – ۱

ISSN: 2745-3673
http://www.Rcsj.ir

مقدمه

سیستم‌های توصیه‌گر به طور کلی به سه دسته‌ی اصلی تقسیم می‌شوند[1]. در رایج‌ترین تقسیم‌بندی، آنها را به سه گروه ۱.محتوا محور ۲.دانش محور و ۳. فیلترینگ تجمعی، تقسیم می‌کنند، که البته گونه چهارمی تحت عنوان Hybrid RS هم برای آنها قائل می‌شوند.

یک رویکرد به سیستم‌های توصیه‌گر، استفاده از الگوریتم‌های CF یا فیلترینگ تجمعی است. در این رویکرد به جای استفاده از محتوای اقلام، از نظرات و رتبه‌بندی‌های انجام شده توسط کاربران برای ارائه پیشنهاد، استفاده می‌شود. لازم به ذکر است که به طور کلی سیستم های مبتنی بر فیلترینگ نیز خود به دو دسته‌ی کاربر محور و کالا محور تقسیم می شوند. همچنین در یک دسته بندی دیگر نیز الگوریتم های CF به دو نوع اصلی مبتنی بر حافظه یا مبتنی بر هیوریستیک و مبتنی بر مدل تقسیم بندی می شوند.

و اما در روش محتوا محور، اقلام پیشنهادی، به این دلیل که با اقلامی که کاربر فعال (کاربری که قرار است به او توصیه کنیم) نسبت به آنها ابراز علاقه کرده است شباهت‌هایی دارند، به کاربر توصیه می‌شوند ولی در CF، لیست اقلام پیشنهادی، بر اساس این اصل که، کاربرانی، مشابه کاربر فعال، از آنها رضایت داشته‌اند تهیه می‌شود. از این رو واضح است که در روش محتوا محور، تمرکز بر روی یافتن شباهت بین اقلام بوده، در حالی که در CF، تمرکز روی یافتن شباهت بین کاربران است؛ بدین ترتیب که پیشنهادات در CF، بر اساس تشابه رفتاری کاربر فعال با کاربران دیگر صورت می‌گیرد و نه بر اساس تشابه ویژگی کالاهای پیشنهادی با ویژگی‌های کالاهای مورد علاقه وی (کاربر فعال).

اما گونهٔ سوم این سیستم‌ها را با نام سیستم‌های دانش محور می‌شناسند. این سیستم‌ها براساس ادراکی که از نیازهای مشتری و ویژگی‌های کالاها پیدا کرده‌اند، توصیه‌هایی را ارائه می‌دهند. به عبارتی در این گونه از سیستم‌های توصیه‌گر، مواد اولیه مورد استفاده برای تولید لیستی از پیشنهادها، دانش سیستم در مورد مشتری و کالا است. سیستم‌های دانش محور از متدهای مختلفی که برای تحلیل دانش، قابل استفاده هستند بهره می‌برند که متدهای رایج در الگوریتم‌های ژنتیک، فازی، شبکه‌های عصبی و ... از جمله آنهاست. همچنین، در این گونه سیستم‌ها از درخت‌های تصمیم، استدلال نمونه‌محور و ... نیز می‌توان استفاده کرد. یکی از رایج‌ترین متدهای تحلیل دانش درسیستم‌های توصیه‌گر دانش محور ،CBR یا روش استدلال نمونه‌محور است.

گونه چهارم، سیستم‌های ترکیبی هستند. طراحان این نوع سیستم‌ها دو یا چند گونه از انواع سه‌گانه مذکور را غالبا به دو منظور با هم ترکیب می‌کنند؛ ۱ -افزایش عملکرد سیستم ۲- کاهش اثر نقاط ضعفی که آن سیستم‌ها وقتی به تنهایی به کار گرفته شوند، دارند. از میان سه روش موجود (CF,CB,KB) ، غالباً روش CF یک پای ثابت این ترکیبات است.

کارایی سیستم‌های توصیه‌گر با توجه به معیارهای مختلفی از جمله دقت، تنوع، خوش اقبالی و قابلیت توسعه‌پذیری ارزیابی می‌شود[2]. بسیاری از روش‌های موجود به دنبال بهبود چنین معیارهایی هستند تا اقلام را به کاربران پیشنهاد دهند. روش‌های توصیه‌گر سنتی مسئله را به صورت تعریف تابع سودمند User×Item→Utility در نظر می‌گرفتند. تابع سودمندی u میزان مفید بودن کالای Itemبرای کاربر را مشخص می‌کند که در آن Utility مجموعه‌ای است که بر اساس میزان اهمیت مرتب شده است [2]. فرایند توصیه معمولا به صورت توصیه K بالاترین آیتم به کاربر با مقدار سودمندی بالا است.

در بسیاری از برنامه‌های کاربردی در نظر گرفتن دو بعد مورد و کاربر کافی نیست چون دیگر اطلاعات برای بهبود کیفیت توصیه مفید هستند. برای مثال ویژگی‌های متنی مثل مکان، زمان و ... یا ویژگی‌های فراداده‌ای مثل سن، جنس و ...عوامل مهمی برای توصیه هستند. . قابلیت مدیریت اطلاعات چندبعدی و انواع مختلفی از توصیه‌ها به عنوان انعطاف پذیری توصیه تعریف می‌شود[3]. در این مقاله برای دستیابی به قابلیت انعطاف پذیری، از رویکرد مبتنی بر گراف استفاده می‌شود. روش توصیه‌ی پیشنهادی بر اساس داده‌های بازخورد ضمنی است. اگرچه داده‌های بازخورد صریح اولویت کاربران را بهتر شرح می‌دهند اما جمع‌آوری این داده‌ها مشکل است. از طرف دیگر داده‌های بازخورد ضمنی حاوی اطلاعات چند بعدی هستند که در این روش مفید است.

۱- کارهای مرتبط

الگوریتم‌های زیادی برای سیستم‌های پیشنهاد دهنده ارایه شده است که هر کدام از آن‌ها ویژگی‌هایی دارند. در روش ما با استفاده از داده‌های بازخورد ضمنی، یک گراف می‌سازیم و سپس از آن جهت توصیه استفاده می‌کنیم، جزء روش پالایش مشارکتی مبتنی بر مدل در نظر گرفته می‌شود.

روش‌های مختلفی جهت دستیابی به انعطاف‌پذیری معرفی شده است[3,6,7,8] . هدف ما در این مقاله دستیابی به انعطاف‌پذیری، با استفاده از مدل کردن روش توصیه‌ی چند بعدی است که بتواند انواع مختلف توصیه‌ها را ارائه دهد.

ISSN: 2745-3673
http://www.Rcsj.ir

چندین روش توصیه‌ی مبتنی بر گراف معرفی شده است[5,9,10,11]. این روش‌ها در دو بعد کار می‌کنند. روش پیشنهادی، ارائه‌ی گراف دوبخشی با استفاده از دامنه‌های مختلف است. الگوریتم رتبه صفحه‌ی شخصی به این گراف اعمال شده و پیشنهادات مناسب ارائه می‌شود.

۲- روش پیشنهادی

همان طور که قبلا گفته شد، مسئله را به صورت دامنه چند بعدی در نظر می‌گیریم. ارائه توصیه به صورت پاسخ دادن به پرس و جوی کاربران انجام می‌شود. روابط بین موجودیت‌ها فراتر از اولویت کاربران روی اقلام است. برای رسیدن به این منظور، یک گراف دوبخشی بر اساس جدول سیاهه[1] داده شده می‌سازیم. برای دستیابی به انعطاف‌پذیری، امکان تعیین دامنه‌ی غیر هدف را به مشتری می‌دهیم. دامنه‌ی غیر هدف را مجموعه‌ی $F=\{f1, f2,\ldots, f_{nf}\}$ در نظر می‌گیریم. w_i را به عنوان وزن f_i تعریف می‌کنیم به طوری‌که داشته باشیم: $w_1+w_2+\ldots+w_{nf}=1$. به دلیل استفاده از دامنه‌ی چند بعدی و در نظر گرفتن مسئله به صورت پاسخ دادن به پرس و جوها اولویت اقلام میان کاربران توسط گام تصادفی[2] مشخص می‌شود.

ماتریس مجاورت وزن دار M برای گراف دوبخشی $G=\{V,E\}$ با اندازه‌ی $|V|\times|V|$ تعریف می‌کنیم. فرض می‌کنیم یال‌های موجود فقط از گره‌هایی که در V_T (دامنه‌ی غیر هدف)نیستند، به گره‌هایی که در V_T (دامنه‌ی هدف) هستند، وصل شوند. با این فرض، گراف G دوبخشی می‌شود.

برای مشخص کردن وزن میان گره‌ها از تعداد سطرهای هم رخداد[3] استفاده می‌کنیم. ماتریس مجاورت M را طوری می‌سازیم که هر عنصر m_{ij} آن متناسب با رابطه‌ی زیر باشد:

$$m_{ij}=\begin{cases} C(i,j) & ,if\ \ v_i\notin V_T\ \ and\ v_j\in V_T \\ C(i,j)\times w_k & ,if\ \ v_i\in V_T, v_j\notin V_T\ and\ v_j\in V_k \\ 0 & ,otherwise. \end{cases} \qquad (1)$$

در این رابطه، $C(i,j)$ تعداد موجودیت‌های هم رخداد متناسب با v_i و v_j است و w_k وزن f_k ای است که متناسب با V_k است. ماتریس احتمال انتقال P را با استفاده از نرمال کردن ماتریس M به صورت زیر می‌سازیم:

$$p_{ij}=\begin{cases} \dfrac{m_{ij}}{\sum_{v_k\in outlink\ [v_i]} m_{ik}} & ,if\ \ outlink[v_i]\neq 0 \\ 0 & ,otherwise. \end{cases} \qquad (2)$$

حال با استفاده از این ماتریس احتمال انتقال، موجودیت‌ها را برای ارائه توصیه‌ی مناسب، رتبه بندی می‌کنیم. الگوریتم رتبه بندی شخصی [12] را برای رتبه‌بندی موجودیت‌ها بر روی مدل گرافی اعمال می‌کنیم. امتیاز رتبه صفحه اصلی به صورت زیر محاسبه می‌شود:

$$\vec{r}=cP^T\vec{r}+(1-c)\frac{1}{n}\vec{e} \qquad (3)$$

n در اینجا تعداد گره‌ها است، r_i میزان امتیاز برای گره v_i است، $\vec{e}=(1,1,\ldots,1)^T$ و $c=0.85$ ضریب ثابت میرایی است. این الگوریتم بر اساس گام تصادفی با امکان راه اندازی مجدد است. برای محاسبه‌ی امتیاز رتبه صفحه‌ی شخصی یک گره، $\frac{1}{n}\vec{e}$ را با بردار اتصال شخصی $\vec{t}$ که علاقه‌مندی کاربران را با بردار بایاس نشان می‌دهد، جایگزین می‌کنیم. در این پژوهش به جای این بردار اتصال از بردار پرس و جوی $\vec{q}$ استفاده می‌کنیم. برای پرس و جوی $Q=\{e_1,e_2,\ldots,e_{n_q}\}$ بردار $\tilde{\vec{q}}$ به صورت زیر تعریف می‌شود:

[1] Log table
[2] Random Walk
[3] Co-occurrences

ISSN: 2745-3673
http://www.Rcsj.ir

$$\tilde{q}_i = \begin{cases} 1 & ,if \ \ \varepsilon_{v_i} \subseteq Q \\ 0 & ,otherwise. \end{cases} \quad (4)$$

ε_{v_i} مجموعه‌ی موجودیت‌هایی است که گره v_i متناسب با آن است. بردار $\tilde{\vec{q}}$ را نرمالیزه کرده و به عنوان بردار اتصال برای رتبه‌بندی موجودیت‌ها در گراف استفاده می‌کنیم. بنابراین رتبه‌بندی گراف با استفاده از $\vec{r} = cP^T \vec{r} + (1-c)\vec{q}$ انجام می‌شود.

نوآوری این مقاله استفاده از اولویت کوتاه مدت[4] است. ثابت شده است که پشت تمایلات کاربر یک زمینه‌ی پویای متنی نهفته است. این تمایلات به صورت مستقیم در گراف نشان داده می‌شود. استدلال این قضیه به این صورت است که اقلامی که در حال حاضر مورد علاقه‌ی کاربر است، تاثیر بیشتری در روند توصیه می‌گذارد. این تاثیر را می‌توان با استفاده از وزن گره‌ها در بردار شخصی q تحت تاثیر قرار داد. به منظور تطبیق K با آخرین کاربر فعال u_a در زمان t از پرس و جو، Q به صورت زیر تعریف می‌شود:

$$Q = \{u_a, s_{t-1}, s_{t-2}, ..., s_{t-k}\} \quad (5)$$

آزمایش بر روی مجموعه‌ی داده‌ی حقیقی (last.fm) استفاده می‌کنیم که شامل ۶۷۳۲۹۳ سیاهه‌ی گوش دادن به موسیقی ۱۶۹۹ کاربران روی ۳۰۰۰ آهنگ می‌باشد.

اندازه‌گیری خطای امتیازدهی معمولاً با استفاده از خطای میانگین مطلق (MAE)[5] یا جذر میانگین مربعات (RMSE)[6] انجام می‌شود. در این روش ما از متریک HR@k استفاده می‌کنیم[8]. این متریک باعث می‌شود بتوان مسئله را به صورت پیش‌بینی موجودیت‌های پنهان در آزمون در نظر گرفت. اگر n را تعداد آزمون‌ها و #hit را تعداد موجودیت‌های پنهان در آزمون در نظر بگیریم، HR@K از رابطه‌ی $\frac{\#hit}{n}$ به دست می‌آید. بدین ترتیب k بالاترین اقلام رتبه بندی شده، در لیست رتبه بندی برای ارائه‌ی توصیه تولید می‌گردد.

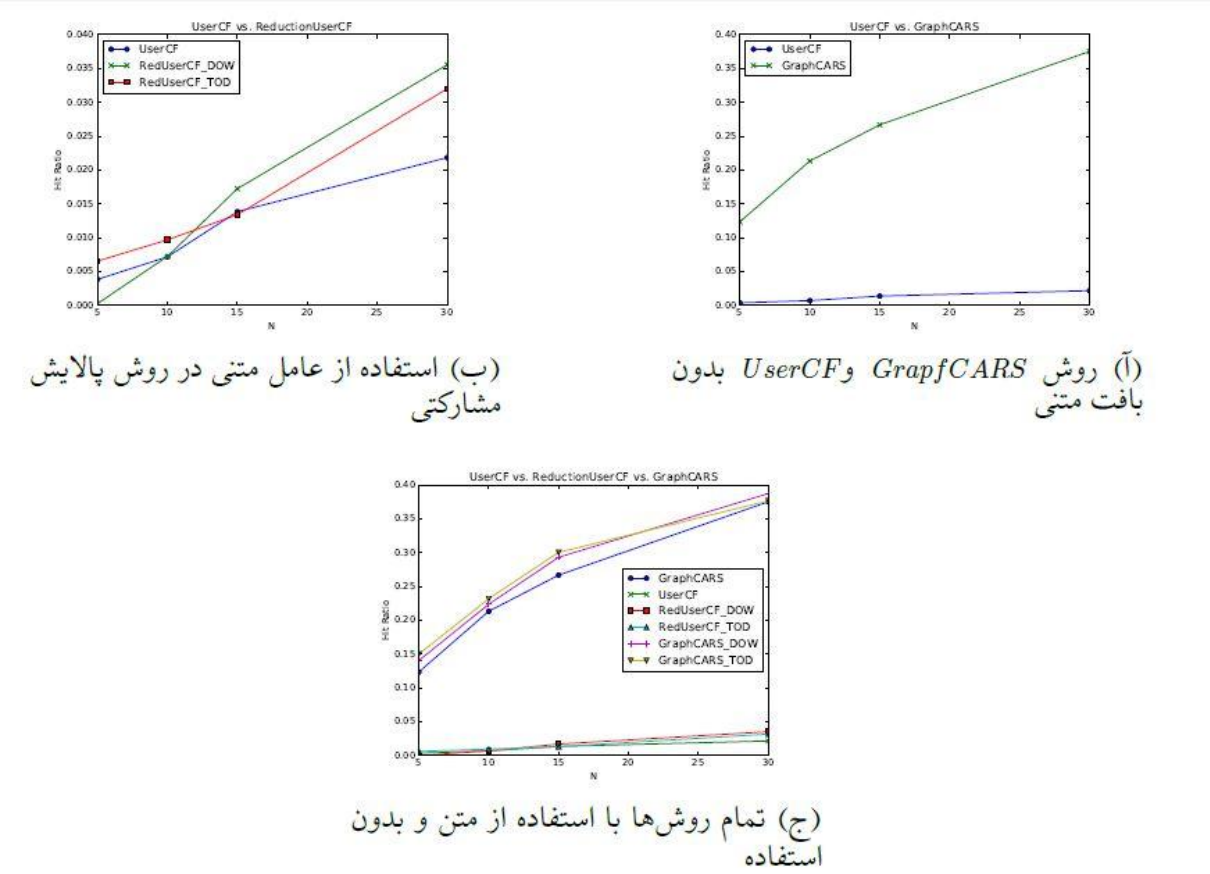
[4] Short-Term Preferences
[5] Mean Absolute Error
[6] Root Mean Square Error

ISSN: 2745-3673
http://www.Rcsj.ir

۳- نتایج

سه آزمایش مختلف انجام گرفته است. هدف از اولین آزمایش مقایسه‌ی رویکرد مبتنی بر گراف با روش‌های پایه‌ی موجود است. یک بار با در نظر گرفتن اطلاعات متنی و بار دیگر بدون در نظر گرفتن اطلاعات متنی. در آزمایش دوم تمرکز بر روی تجزیه و تحلیل اطلاعات و ویژگی‌های متنی در سیستم‌های مبتنی بر گراف است. در هر مورد تنظیمات مختلفی با هم مقایسه می‌شوند. در آخرین آزمایش اولویت‌های کوتاه مدت دقیقاً مورد بررسی قرار گرفته است. بهینه‌سازی در این قسمت آزمایش بیشتر مورد توجه است.

ابتدا روش مبتنی بر گراف را با روش‌های پالایش مشارکتی مقایسه می‌کنیم. در هر دو مورد تعداد همسایه‌ها را k=5 در نظر می‌گیریم. نتیجه‌ی حاصل از این آزمون در شکل ۱ نشان داده شده است.



(آ) روش $GrapfCARS$ و$UserCF$ بدون بافت متنی

(ب) استفاده از عامل متنی در روش پالایش مشارکتی

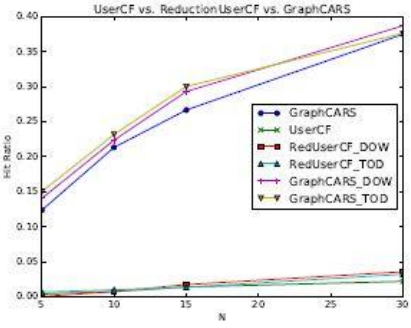
(ج) تمام روش‌ها با استفاده از متن و بدون استفاده

شکل ۱: مقایسه‌ی HitRatio در روش‌های پایه با روش مبتنی بر گراف

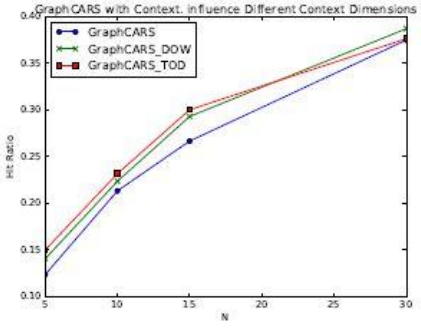
در شکل ۲ نتیجه‌ی مقایسه‌ی HitRatio در ترکیب مختلف متنی نشان داده شده است. در این مجموعه از این آزمون اطلاعات مختلف به نمودار اضافه می‌شود. با توجه به تاثیر این اطلاعات، توصیه ساخته می‌شود. سیستم توصیه‌گر آگاه از متن مبتنی بر گراف، به عنوان روش پایه و بدون هیچ اطلاعات اضافی در نظر گرفته می‌شود. تمامی ضرایب به صورت یکسان وزن می‌گیرند.

نتایج استفاده از اولویت کوتاه مدت در شکل ۳ نشان داده شده است. اندازه‌ی k به صورت قابل توجهی در میزان دقت پیشگوئی تاثیر دارد. بهترین روش متنی هنگامی بدست می‌آید که k=6 انتخاب شود.

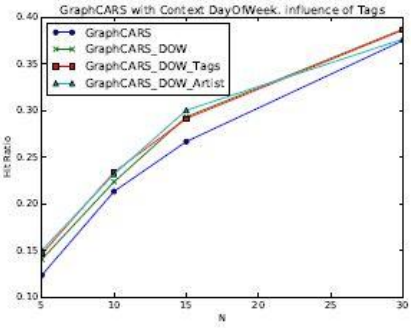
ISSN: 2745-3673
http://www.Rcsj.ir



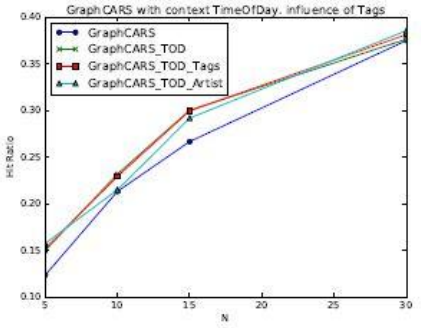
(آ) تاثیر عامل زمان و خواننده بدون بافت متنی



(ب) تاثیر عامل بافت زمانی

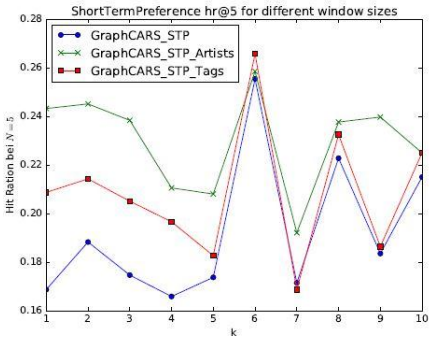


(ج) تاثیر عامل متنی روز هفته و برچسب

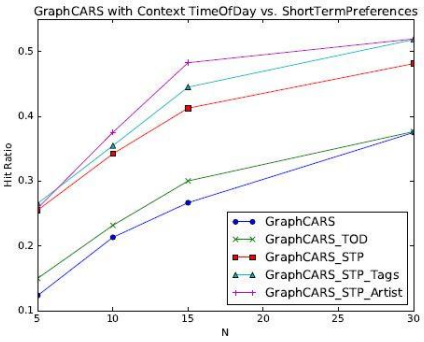


(د) تاثیر عامل متنی ساعت روز و برچسب

شکل ۲: مقایسه‌ی HitRatio با گراف‌های مختلف



(آ) مقایسه‌ی $HR@5$ از $GraphCARS_STP$ با اندازه پنجره‌های مختلف



(ب) مقایسه‌ی $GraphCARS_STP$ با $GraphCARS$ با و بدون متن. اندازه‌ی پنجره: $k = 6$

شکل ۳: تاثیر اولویت‌های کوتاه مدت

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[1] Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." IEEE transactions on knowledge and data engineering 17.6 (2005): 734-749.Williams, A., "Experimental investigation of premixed combustion within highly porous media", Proceeding of the ASME/JSME Thermal Engineering Joint Conference, pp. 752-758, 1992.

[2] Steck, Harald. "Training and testing of recommender systems on data missing not at random." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[3] Adomavicius, Gediminas, Alexander Tuzhilin, and Rong Zheng. "REQUEST: A query language for customizing recommendations." Information Systems Research 22.1 (2011): 99-117.

[4] Lee, Dongjoo, et al. "Exploiting contextual information from event logs for personalized recommendation." Computer and Information Science 2010. Springer Berlin Heidelberg, 2010. 121-139.

[5] Gori, Marco, et al. "ItemRank: A Random-Walk Based Scoring Algorithm for Recommender Engines." IJCAI. Vol. 7. 2007.

[6] Koutrika, Georgia, Benjamin Bercovitz, and Hector Garcia-Molina. "FlexRecs: expressing and combining flexible recommendations." Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. ACM, 2009.

[7] Adomavicius, Gediminas, et al. "Incorporating contextual information in recommender systems using a multidimensional approach." ACM Transactions on Information Systems (TOIS) 23.1 (2005): 103-145.

[8] Kahng, Minsuk, Sangkeun Lee, and Sang-goo Lee. "Ranking in context-aware recommender systems." Proceedings of the 20th international conference companion on World wide web. ACM, 2011.

[9] Baluja, Shumeet, et al. "Video suggestion and discovery for youtube: taking random walks through the view graph." Proceedings of the 17th international conference on World Wide Web. ACM, 2008.

[10] Desrosiers, Christian, and George Karypis. "A novel approach to compute similarities and its application to item recommendation." Pacific Rim International Conference on Artificial Intelligence. Springer Berlin Heidelberg, 2010.

[11] Cheng, Haibin, et al. "Recommendation via query centered random walk on k-partite graph." Seventh IEEE International Conference on Data Mining (ICDM 2007). IEEE, 2007.

[12] Basch, Julien, Leonidas J. Guibas, and John Hershberger. "Data structures for mobile data." Journal of Algorithms 31.1 (1999): 1-28.