مطالعات علوم کاربردی در مهندسی

دوره ۸، شماره ۳، پاییز ۱۴۰۱

صفحات ۸۴-۶۸

Online ISSN: ۲۴۷۶-۴۵۰۷

Print ISSN: ۲۵۳۸-۲۰۵۵

www.irijournals.com

پیمایش هوشمند وب با کمک تحلیل پیوندی سند و مدل مارکوف

الهام صادقی[۱]، الهام کاظمی[۲]

[۱] کارشناسی ارشد مهندسی کامپیوتر دانشگاه آزاد اسلامی واحد آبادان

[۲] استاد راهنما و عضو عیئت علمی دانشگاه آزاد اسلامی واحد آبادان

چکیده

با توجه به افزایش حجم وب، پیمایش و جستجوی آن از اهمیت بالایی برخوردار است. در پیمایش این حجم وسیع از صفحات بهتر آن است، صفحاتی ابتدا پیمایش شوند که مرتبط با موضوع مورد نظر می‌باشند. "پیمایش هوشمند" روشی است که برای پیمایش صفحات مرتبط با یک موضوع به کار می‌رود. در این روش سعی بر آن است که در هنگام پیمایش، صفحات hub (صفحاتی ست که به صفحات مهمی اشاره می‌کنند) خوب تشخیص داده شوند تا از آن‌ها به عنوان منبعی برای رسیدن به صفحات authoriry (صفحاتی هستند که محتوای مهمی دارند) استفاده شود. در این تحقیق از "الگوریتم Page Rank" که یکی از الگوریتم مبتنی بر مدل‌های مارکوف است استفاده می‌شود و همچنین از تکنیک‌ها وب کاوی نظیر "تحلیل پیوند" و "کاوش استفاده از وب" برای ساخت پروفایل کاربران استفاده می‌شود. اساس کار اینگونه است که برای استخراج الگوهای حرکتی با استفاده از معیارهای "مدت زمان مشاهده صفحه" و "فرکانس مشاهده صفحه" که به خوبی میزان اهمیت و علاقه کاربران به آن صفحه را نشان می‌دهد وزن صفحات را محاسبه می‌کند. برای پیدا کردن صفحاتی که بیشتر "مدت زمان مشاهده صفحه" و "فرکانس مشاهده صفحه" را دارد از الگوریتم Gradual Extra Weight (GEW) استفاده می‌شود. سپس برای حل مشکل کیفیت پایین پیشنهاد بیش از یک صفحه، از معیار "تحلیل پیوند صفحات" از الگوریتم (c3) Contextual Concept Clustering برای خوشه بندی استفاده شده است. در نهایت با استفاده ازالگوریتم page rank که لیستی از صفحات دلخواه کاربر را به وی پیشنهاد می‌دهد.

واژه‌های کلیدی: پیمایش هوشمند، تحلیل پیوندی، الگوریتم c3، الگوریتم GEW، مدل مارکوف.

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۸۴-۶۸

۱- مقدمه

با توسعه سیستم‌های اطلاعاتی، داده به یکی از منابع پر اهمیت سازمان‌ها مبدل گشته است. بنابراین روش‌ها و تکنیک‌هایی برای دستیابی کارا به داده، اشتراک داده، استخراج اطلاعات از داده و استفاده از این اطلاعات، مورد نیاز می‌باشد. با ایجاد و گسترش وب و افزایش چشمگیر حجم اطلاعات، نیاز به این روش‌ها و تکنیک‌ها بیش از پیش احساس می‌شود. وب، محیطی وسیع، متنوع و پویا است که کاربران متعدد اسناد خود را در آن منتشر می‌کنند. در حال حاضر بیش از دو میلیون صفحه در وب موجود است و این تعداد با نرخ ۷. ۳ میلیون صفحه در روز افزایش می‌یابد. با توجه به حجم وسیع اطلاعات در وب، مدیریت آن با ابزارهای سنتی تقریبا غیر ممکن است و ابزارها و روش‌هایی نو برای مدیریت آن مورد نیاز است.

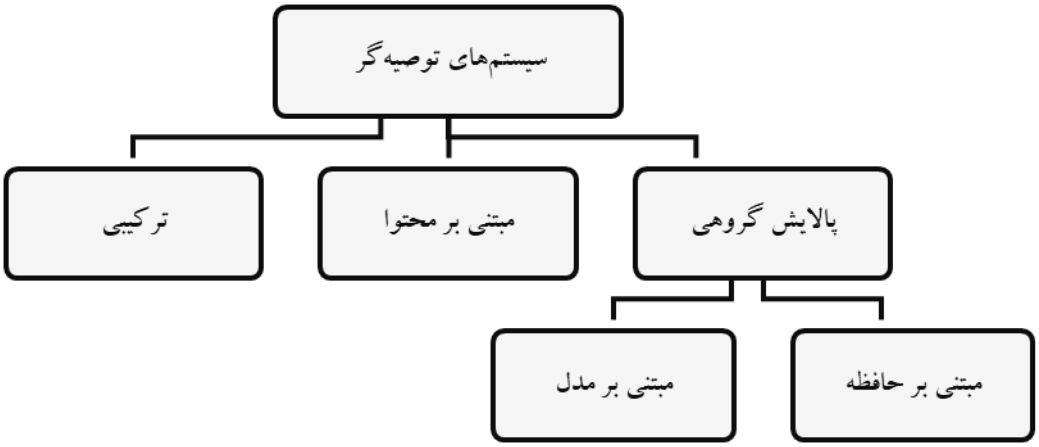
وب طی یک فرآیند آشفته و غیر متمرکز در حال رشد است و این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است که از هیچ گونه سازماندهی منطقی برخوردار نیستند. در واقع وب به مجموعه بزرگی از داده‌های ساخت یافته و نیمه ساخت یافته تبدیل شده است که کاربران آن از همپوشانی داده‌ها رنج می‌برند. بنابراین تحلیل رفتارهای کاوشی کاربران وب و بررسی واقعی علایق کاربران اهمیت خاصی پیدا کرده است. بررسی رفتارهای کاربران در وب، به عنوان روشی جهت کشف دانش نهفته در نحوه تعامل کاربران با وب، یکی از ابزارهای مهم در حوزه کاوش در وب شناخته می‌شود. با توجه به افزایش حجم وب، پیمایش و جستجوی آن از اهمیت بالایی برخوردار است. در پیمایش این حجم وسیع از صفحات بهتر آن است، صفحاتی ابتدا پیمایش شوند که مرتبط با موضوع مورد نظر می‌باشند. "پیمایش هوشمند" روشی است که برای پیمایش صفحات مرتبط با یک موضوع به کار می‌رود. در این روش سعی بر آن است که در هنگام پیمایش، صفحات hub (صفحاتی ست که به صفحات مهمی اشاره می‌کنند) خوب تشخیص داده شوند تا از آن‌ها به عنوان منبعی برای رسیدن به صفحات authoriry (صفحاتی هستند که محتوای مهمی دارند) استفاده شود. [۲]

وب طی یک فرآیند آشفته و غیر متمرکز در حال رشد است و این روند منجر به تولید حجم وسیعی از مستندات متصل به یکدیگر گشته است که از هیچ گونه سازماندهی منطقی برخوردار نیستند. در واقع وب به مجموعه بزرگی از داده‌های ساخت یافته و نیمه ساخت یافته تبدیل شده است که کاربران آن از همپوشانی داده‌ها رنج می‌برند. بنابراین تحلیل رفتارهای کاوشی کاربران وب و بررسی واقعی علایق کاربران اهمیت خاصی پیدا کرده است. بررسی رفتارهای کاربران در وب، به عنوان روشی جهت کشف دانش نهفته در نحوه تعامل کاربران با وب، یکی از ابزارهای مهم در حوزه کاوش در وب شناخته می‌شود. کارهای تحقیقاتی بسیاری در این حوزه انجام شده است که عمدتا بر مبنای اطلاعات موجود از رفتار کاربر در تعامل با وب به استخراج این دانش و استفاده از آن در کاربردهای مختلف در وب، نظیر شخصی سازی صفحات وب خود سازمانده کردن وب می‌پردازند. اما سیستم‌های که برای ارائه پیشنهادات فقط از رفتار کاربران استفاده می‌کنندبه دلیل مشکلات زیر، عموماً از دقت پایینی برخوردار هستند و ممکن است صفحات با ارزشی در بخش پیشنهاد صفحات فراموش شوند.

هدف این مقاله ، با استفاده از تکنیک‌های وب کاوی ، از جمله پیمایش هوشمند در ساختار وب و محتوای وب و با استفاده از الگوریتم رتبه بندی page rank به حل مشکل عدم دسترسی آسان به اطلاعات مورد نیاز در زمان مناسب می‌پردازد و همچنین مشکل کاهش دقت الگوریتم‌ها با افزایش تعداد صفحات پیشنهادی را نیز بررسی می‌کند. با توجه به مطالب بالا، آنچه موجب نوآوری این تحقیق می‌باشد. انتخاب الگوریتم‌های ترکیبی Gradual Extra Weight(GEW) و (c3) Contextual Concept Clustering به منظور پیشنهاد صفحات به کاربران می‌باشد. که انتخاب این الگوریتم‌ها باعث شده دقت در پیمایش افزایش یابد.

دسته بندی سیستم‌های توصیه‌گر

سیستم‌های توصیه‌گر سنتی را می‌توان با توجه به آنچه در شکل (۱) نشان داده شده است، به صورت زیر دسته‌بندی نمود، که در ادامه توضیحات بیشتر را در مورد هر یک ارائه خواهیم کرد.



شکل (۱) ساختار سلسله مراتبی از سیستم‌های توصیه‌گر [۵۰]

روش پیشنهادی

در این بخش به بررسی روش پیشنهادی بر اساس دو الگوریتم ۳c,GEW بر اساس [۴۲] پرداخته خواهد شد و سپس با استفاده از مدل مارکوف به عنوان ایده‌ای جدید به ترکیب این دو الگوریتم و مدل مارکوف پرداخته خواهد شد.

ویژگی‌ها ومدل سازی پویای پروفایل کاربران

قبل از پرداختن به ارائه روش پیشنهادی باید دید که یک پروفیل خوب و پویا چه ویژگی‌هایی دارد. برای این منظور در این بخش به ارائه ویژگی‌ها لازم در این زمینه پرداخته خواهد شد:

کاربران تمایلی به ارائه اطلاعات در مورد علایق و اولویت‌های خود به صراحت از طریق تکمیل پرسشنامه و . . . ندارند از این رو، یک سیستم شخصی سازی انعطاف پذیر باید تلاش کند از طریق ردیابی رفتارهای کاربر به طور ضمنی بدون دخالت کاربر و به همان اندازه دقیق علایق و اولویت‌های کاربر را بدست آورد.

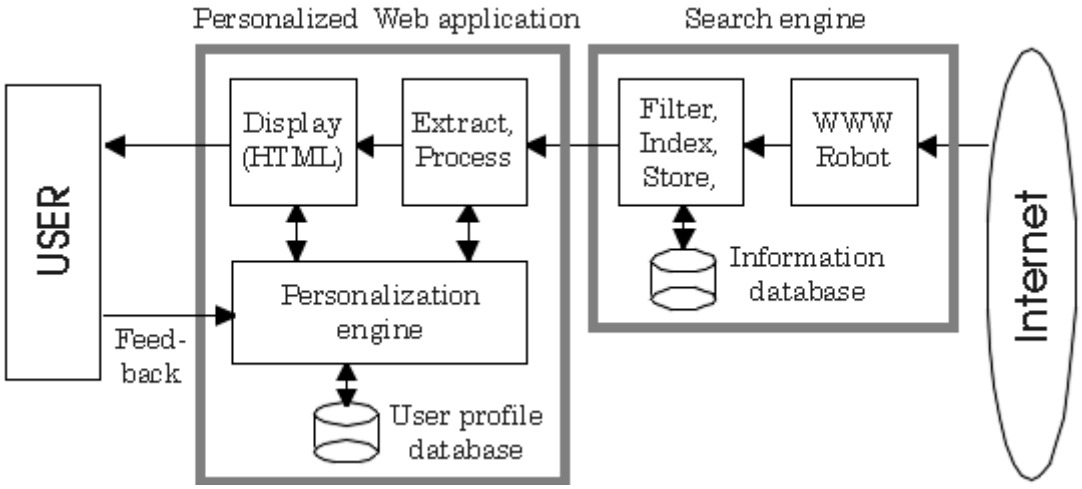
به منظور ارائه خدمات پیشرفته به کاربران، باید علایق و اولویت‌های آن‌ها در یک قالب مناسب ذخیره شوند تا پردازش بیشتری انجام شود و اطلاعات مفیدی را بتوان استخراج کرد. برای این منظور، از هستی‌شناسی برای نمایش علایق کاربر به منظور پشتیبانی موثرتر از لحاظ معنایی از پروفایل و در نتیجه توصیه‌های غنی‌تر استفاده شود.

علایق کاربران بندرت ثابت باقی می‌ماند و به طور مداوم در طول زمان تغییر می‌کند و تکامل می‌یابد. اکثر سیستم‌های شخصی سازی موجود با این چالش در ثبت رفتار کاربر دست و پنجه نرم نمی‌کنند و یا این کار را به طریقی انجام نمی‌دهند که به شیوه‌ای جامع به رفتار کاربر بپردازند. هدف این پژوهش توسعه روش‌های پروفایل است که بتواند قادر به ردیابی تغییر در علایق کاربر و سازگاری با آن باشد.

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۸۴-۶۸

علایق کاربر می‌تواند در یک دوره کوتاه بسیار متمایز باشد، که شامل علایق فعلی کاربر باشد و می‌تواند بسیار متغییر و بلند مدت نیز باشد و میل بیشتری به پایداری در طول زمان دارند. چنین علایق و رفتار متمایزی از کاربر می‌تواند برای ذخیره بسیار مفید باشد . این ذخیره اطلاعات می‌تواند در درتلاش برای درک هر چه بیشتر نیازهای کاربر به صورت منحصر به فرد و در ابعاد بیشتر یاری کند. این علایق متمایز در اغلب تحقیقات نادیده گرفته شده است.

به طور معمول در شخصی سازی کار برای حل مشکلات خاص دامنه توسعه داده شده است. هدف از این پژوهش توسعه مدل‌ها و روش‌هایی است که کاربردهای وسیع تری دارند و می‌توانند یکپارچه و مستقر شوند تا بتوانند خدمات متنوعی را ارائه دهند. در شکل (۲) می‌توان یک فرآیند شخصی سازی و ایجاد پروفایل برای کاربران را مشاهده نمود.



شکل (۲) فرآیند ساخت پروفایل [۴۲]

اما این داده‌ها و علایق رسیده از کاربر احتیاج به یک سری پالایش داده دارد تا بوسیله الگوریتم ارائه شده در [۴۲] و مدل مارکوف ارائه شده در این پژوهش به تحلیل رفتار کاربر پرداخت.

مدل مارکوف بر اساس همه مراتب

رفتار پیمایشی کاربر بر روی سایت بوسیله بررسی صفحاتی که مشاهده کرده است مدل می‌شود این مجموعه از صفحه‌ها به عنوان یک دوره (w) در نظر گرفته می‌شود و به صورت دنباله ای از صفحات که توسط کاربر بازدید شده‌اند نمایش داده می‌شود. در این پژوهش نیز این مدل به عنوان یک مدل پیش بینی در کنار الگوریتم ۳c مورد استفاده قرار می‌گیرد تا نتایج این الگوریتم را بهبود دهد. مسئله پیش بینی صفحه بعد می‌تواند توسط یک روش احتمالی به این صورت حل شود فرض کنید w دوره وب کاربر با طول n است. و $P(P_i|W)$ احتمال اینکه صفحه P_i به عنوان صفحه بعد بازدید شود می‌باشد. سپس صفحه P_{n+1} از فرمول زیر محاسبه می‌شود.

$$P_{n+1} = \arg\max_{p\in\rho} \{P(p_{n+1} = p \mid w)\} = \arg\max_{p\in\rho} \{P(p_{n+1} = p \mid p_n, p_{n-1}, p_{n-2}, \ldots, p_1)\}$$

(ρ مجموعه کل صفحات موجود بر روی وب سایت است)

در اصل با این روش احتمال تمام صفحات برای صفحه بعدی بودن محاسبه و سپس صفحه‌ای با بیشترین احتمال به عنوان پیش بینی انتخاب می‌شود. اما به دلیل اینکه محاسبه تمام این احتمالات شرطی به صورت دقیق غیر ممکن می‌باشد از فرآیند

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۸۴-۶۸

مارکوف برای پیش بینی صفحه بعد استفاده می‌شود که با توجه به این فرآیند تنها آخرین k صفحه ی دیده شده توسط کاربر برای پیش بینی استفاده می‌شود تعداد صفحات k مشخص کننده مرتبه مدل مارکوف می‌باشدو فرمول مشخص کردن صفحه P_{n+1} بصورت زیر تبدیل می‌شود:

$$P_{n+1} = \arg\max_{p\in\rho} \{P\,(\,p_{n+1} = p \mid p_n, p_{n-1}, p_{n-2}, \ldots, p_{n-(k-1)}\,)\}$$

مدل مارکوف مورد استفاده در این پژوهش مدل مارکوف با همه مراتب می‌باشد.

مدل مارکوف با همه مراتب

در بسیاری از موارد مدل‌های مارکوف با مرتبه پائین (اول یا دوم) قادر به پیش بینی دقیق صفحه بعدی که توسط کاربر دیده خواهد شد نیستند و این امر بدین دلیل است که این مدل‌ها به نگاهی عمیقی در گذشته کاربر نمی‌پردازد و تنها براساس مشاهده یک یا دو صفحه آخر پیش بینی می‌کنند و در نتیجه برای گرفتن دقت بهتر باید از مدل‌های مارکوف با مرتبه بالاتر (سه یا چهار) استفاده کرد. متاسفانه مدل‌های مارکوف مرتبه بالاتر نیز محدودیت‌هایی دارند از جمله تعداد بالای حالت‌ها، پوشش کم و حتی گاهی بدلیل پوشش کم دقت کمتری دارند. یک راه برای حل مشکل پوشش کم یادگیری انواع مرتبه‌های مدل مارکوف و سپس ترکیب آن‌ها برای پیش بینی می‌باشد. در این رویه برای هر نمونه از بزرگترین مرتبه مدل مارکوفی که نمونه را پوشش دهد برای پیش بینی استفاده می‌شود. برای مثال اگر این مدل شامل سه مرتبه از مدل مارکوف باشد نمونه داده شده اول در صورت امکان با مرتبه سه پیش بینی می‌شود و اگر توسط مرتبه سه پوشش داده نشود این روال برای مرتبه دو و یک هم تکرار می‌شود. این روش مدل مارکوف با همه مراتب نام دارد این روش مشکل پوشش را به خوبی حل می‌کند و می‌تواند دقت دسته بندی‌های الگوریتم ۳C را تا حد بسیار بالای بهبود ببخشد.

مدل پنهان مارکوف

یک مدل مارکوف آماری است که در آن سیستم مدل شده به صورت یک فرایند مارکوف با حالت‌های مشاهده نشده (پنهان) فرض می‌شود. یک مدل پنهان مارکوف می‌تواند به عنوان ساده‌ترین شبکه بیزی پویا در نظر گرفته شود. در مدل عادی مارکوف، حالت به طور مستقیم توسط ناظر قابل مشاهده است و بنابراین احتمال‌های انتقال بین حالت‌ها تنها پارامترهای موجود است. در یک مدل پنهان مارکوف، حالت به‌طور مستقیم قابل مشاهده نیست، اما خروجی، بسته به حالت، قابل مشاهده است. توجه داشته باشید که صفت "پنهان" به دنباله حالت‌هایی که مدل از آن‌ها عبور می‌کند اشاره دارد، نه به پارامترهای مدل؛ حتی اگر پارامترهای مدل به‌طور دقیق مشخص باشند، مدل همچنان "پنهان" است.

محیط شبیه سازی

MATLAB یک زبان سطح بالا و با محیطی جذاب می‌باشد، که در ابتدا براساس زبان برنامه نویسی C توسعه داده شد. واژه متلب هم به معنی محیط محاسبات رقمی و هم به معنی خود زبان برنامه‌نویسی مربوطه‌است که از ترکیب دو واژه MATrix (ماتریس) وLABoratory (آزمایشگاه) ایجاد شده‌است. این نام حاکی از رویکرد ماتریس محور برنامه است، که در آن حتی اعداد منفرد هم به عنوان ماتریس در نظر گرفته می‌شوند در این پژوهش برای شبیه سازی سیستم از نرم افزار matlab

استفاده شده است زیرا بوسیله این نرم افزار می‌توان داده‌های ورودی را در غالب ماتریس‌هایی تعریف نمود و به مدل سازی سیستم پرداخت.

جامعه آماری و حجم نمونه

در این سیستم از داده‌های جمع‌آوری شده بر اساس دوازده موضوع انتخابی تصادفی که در مدت بیست روز جمع‌آوری شده است استفاده خواهد شد. این داده‌ها براساس مقاله [۴۲] نشان دهنده رفتار هر کاربر براساس پنج سناریو می‌باشد. نکته مهم راجب این موضوعات تقسیم بندی موضوعات در سه نوع می‌باشد که در جدول (۱) نمایش داده شده است.

جدول (۱) نوع علاقه به موضوع

شرح	نوع
موضوعه با علاقه زیاد	long
موضوعه با علاقه کم	Short
عدم وجود علاقه به موضوع	unintersting

سناریو اول (رفتار نرمال کاربر)

در سناریوی اول، نگاهی به نحوه یادگیری و سازگاری روش پیشنهادی با رفتار عادی کاربر انداخته می‌شود. این سناریو رفتار کاربر را در زمان وب گردی عادی مورد تحلیل قرار می‌دهد. و میزان این رفتار عادی در جدول (۲) در هر روز بر اساس مقادیر صفر تا چهار نمایش داده شده است.

جدول (۲) رفتار کاربر بر اساس سناریو اول در مدت بیست روز [۴۲]

Day / Topic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Type
Topic 1	0	1	2	3	1	3	2	0	2	1	1	2	2	0	0	1	2	3	2		Long
Topic 2	2	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	2	0	Uninteresting
Topic 3	2	3	4	2	1	2	0	1	1	0	0	1	1	0	1	2	2	0	1	1	Long
Topic 4	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	Uninteresting
Topic 5	4	3	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Short
Topic 6	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	Uninteresting
Topic 7	0	0	0	0	0	0	0	0	0	0	0	0	1	2	2	4	1	2	2	1	Short

Topic 8	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	Uninteresting
Topic 9	0	0	0	0	0	0	0	0	0	0	1	1	2	3	2	2	2	1	0	0	Short
Topic 10	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	Uninteresting
Topic 11	2	1	2	3	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	Short
Topic 12	0	2	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	Uninteresting

سناریو دوم (توقف عملیات وب گردی)

در این سناریو کاربر به طور ناگهانی وموقت جستجوی خود را متوقف می‌کند و دوباره به جستجوی خود ادامه می‌دهد. که هر چقدر اعداد جدول (۳) بیشتر باشد مدت طولانی تری مرور گر راجب آن موضوع توقف پیدا کرده است.

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۸۴-۶۸

جدول (۳) رفتار کاربر بر اساس سناریو دوم در مدت بیست روز [۴۲]

Day / Topic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Type
Topic 1	0	2	3	1	1	1	1	0	0	0	0	0	1	2	3	4	3	2	2	1	Long
Topic 2	2	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	Uninteresting
Topic 3	3	3	4	4	3	2	1	0	0	0	0	0	1	2	2	2	2	3	2	2	Long
Topic 4	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	Uninteresting
Topic 5	2	2	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Short
Topic 6	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	Uninteresting
Topic 7	0	0	0	0	3	2	2	0	0	0	0	0	2	3	3	1	1	0	0	0	Short
Topic 8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Uninteresting
Topic 9	0	0	0	0	0	0	0	0	0	0	0	0	1	2	3	1	1	0	0	0	Short
Topic 10	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2	0	Uninteresting
Topic 11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	4	3	3	1	1	Short
Topic 12	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	Uninteresting

سناریو سوم (میزان ریسک تغییر رفتار ناگهانی کاربر راجب یک موضوع)

در این سناریو میزان ریسک تغییر ناگهانی رفتار کاربر در سیستم بازیابی شده است که البته در این مسئله روز وموضوع مورد بررسی کاربر بسیار تعیین کننده می‌باشند. این شدت میزان ریسک با اعداد صفر تا چهار نمایش داده می‌شود. در جدول (۴) مقادیر بر اساس موضوعات و روزها نمایش داده شده است.

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۸۴-۶۸

جدول (۴) رفتار کاربر بر اساس سناریو سوم در مدت بیست روز [۴۲]

Day / Topic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Type
Topic 1	0	0	1	2	3	3	2	0	1	0	0	0	1	2	4	2	1	1	0	0	Long
Topic 2	2	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	2	0	Uninteresting
Topic 3	2	3	4	3	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	Short
Topic 4	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	Uninteresting
Topic 5	0	0	0	0	0	0	2	4	4	4	3	1	1	1	0	0	0	0	0	0	Short
Topic 6	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	Uninteresting
Topic 7	0	0	0	0	0	0	4	4	3	3	2	1	1	0	0	0	0	0	0	0	Short
Topic 8	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	Uninteresting
Topic 9	0	0	0	0	0	0	0	0	0	0	1	1	2	3	2	2	2	1	1	1	Short
Topic 10	0	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	Uninteresting
Topic 11	1	2	3	3	4	1	2	1	0	0	3	4	2	2	3	2	1	1	2	1	Long
Topic 12	0	1	1	2	3	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	Short

سناریو چهارم (تغییر رفتار ناگهانی کاربر از یک موضوع به موضوع دیگر)

در این سناریو کاربر به طور ناگهانی علاقه خود را از یک موضوع به موضوع دیگر معطوف می‌کند که شدت تغییر رفتار با اعداد صفر تا چهار نمایش داده می‌شود. در جدول (۵) مقادیر براساس موضوعات و روزها نمایش داده شده است.

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۸۴-۶۸

جدول (۵) رفتار کاربر بر اساس سناریو چهارم در مدت بیست روز [۴۲]

Day / Topic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Type
Topic 1	0	0	1	2	3	3	2	0	1	0	0	0	1	2	4	2	1	1	0	0	Long
Topic 2	2	0	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	2	0	Uninteresting
Topic 3	2	3	4	3	3	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	Short
Topic 4	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	Uninteresting
Topic 5	0	0	0	0	0	0	2	4	4	4	3	1	1	1	0	0	0	0	0	0	Short
Topic 6	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	Uninteresting
Topic 7	0	0	0	0	0	0	4	4	3	3	2	1	1	0	0	0	0	0	0	0	Short
Topic 8	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	Uninteresting
Topic 9	0	0	0	0	0	0	0	0	0	0	1	1	2	3	2	2	2	1	1	1	Short
Topic 10	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	Uninteresting
Topic 11	1	2	3	3	4	1	2	1	0	0	3	4	2	2	3	2	1	1	2	1	Long
Topic 12	0	1	1	2	3	3	2	1	0	0	0	0	0	0	0	0	0	0	0	0	Short

سناریو پنجم (وقفه در رفتار عادی کاربر)

در این سناریو کاربر برای مدت کوتاهی در رفتار خود تغییر می‌دهد به عنوان مثال رفتار روتین کاربر مطالعه مقالات علمی و جستجو ویافتن آن‌ها است ولی به علت تعطیلات و مسافرت به کشور دیگری موضوعات وب گردی خود را راجب جاذبه‌های توریستی کشور مورد نظر معطوف می‌کند که این تغییر رفتار موقتی و با یک وقفه کوتاه می‌باشد. در جدول (۶) رفتار کاربر بر اساس سناریو مطرح شده نمایش داده شده است.

جدول (۶) رفتار کاربر بر اساس سناریو پنجم در مدت بیست روز [۴۲]

Day / Topic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Type
Topic 1	4	4	3	3	3	2	0	0	0	0	0	1	1	2	3	2	2	2	1	1	Long
Topic 2	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	Uninteresting
Topic 3	2	3	4	2	3	2	0	0	0	0	0	2	0	1	3	2	1	0	1	1	Long
Topic 4	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	Uninteresting
Topic 5	0	0	0	0	0	0	4	3	3	1	2	0	0	0	0	0	0	0	0	0	Short(Sudden)
Topic 6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	Uninteresting
Topic 7	0	0	0	0	0	0	3	3	3	3	2	0	0	0	0	0	0	0	0	0	Short(Sudden)
Topic 8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Uninteresting
Topic 9	0	0	1	2	3	2	0	0	0	0	0	2	2	1	1	2	0	0	0	0	Short
Topic 10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	Uninteresting
Topic 11	0	0	0	0	4	3	0	0	0	0	0	2	1	3	2	1	0	0	0	0	Short
Topic 12	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	Uninteresting

ارزیابی روش پیشنهادی

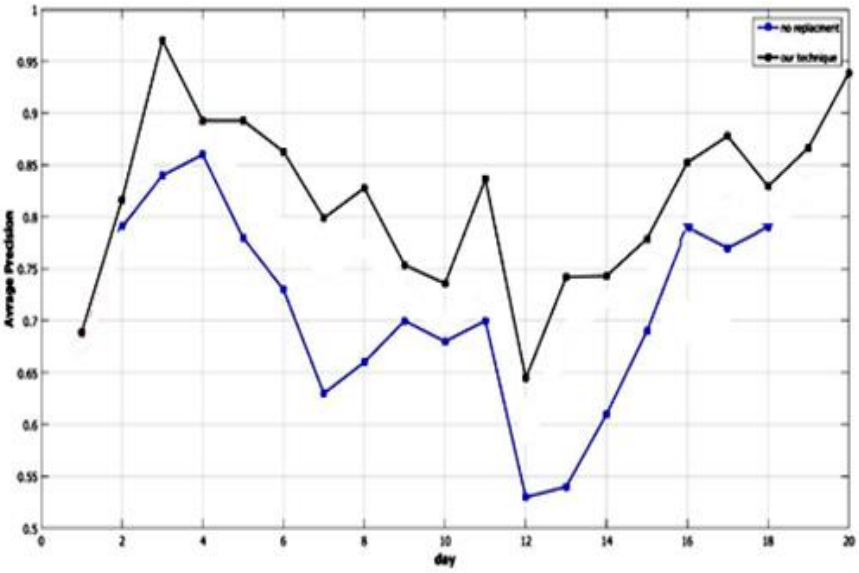
در اینجا هدف ما ارزیابی عملکرد نمایه کاربری پویا پیشنهاد شده در این پژوهش می‌باشد و اثربخشی فرآیندهای یادگیری و تطبیق است. از این رو، ابتدا ویژگی‌های مختلف سیستم مورد ارزیابی قرار می‌گیرند و سپس عملکرد کلی روش پیشنهادی در مقایسه با سایر روش‌های مدل سازی مورد مقایسه قرار می‌گیرند. برای همه این ارزیابی‌ها، دقت یادگیری و سازگاری پروفایل کاربر را برای هر روز از زمان آزمایش (۲۰ روز) با استفاده از رابطه زیر آزمایش می‌شود.

دوره ۸، شماره ۳، پاییز ۱۴۰۱، صفحات ۸۴-۶۸

$$Precision = \frac{|Number\ of\ correct\ learned\ and\ adapted\ interests|}{|Total\ number\ of\ actual\ interests|}$$

براساس این تعداد تشخیص صفحات مورد علاقه کاربر توسط سیستم به کل صفحات مورد علاقه کاربر به عنوان مقدار دقت در نظر گرفته شده است. دقت در این معادله به عنوان نسبت موضوعات مورد علاقه به درستی مدل شده در پروفایل‌های کاربر به تعداد واقعی موضوعات مورد علاقه در سناریوها می‌باشد.

در این بخش به مقایسه روش ارائه شده در این پژوهش با روش ارائه شده در مقاله [۴۲] با نام (re-ranking model) پرداخته می‌شود در این مقایسه به این نکته پرداخته خواهد شد که اگر سیستم پیشنهادی صفحات وب برای کاربر فعال نباشد تا چه حد به صفحات مورد درخواست خود به چند درصد خطا می‌رسد که در نمودار (۱) با نام (no-replacment) و زمانی که روش re-ranking و روش مدل مارکوف در این پژوهش برای کاربر فعال باشد دقت سیستم تاچه حد تغییر می‌یابد. این آزمایشات برای بیست روز انجام شده است .



نمودار (۱) مقایسه زمانیکه سیستم توصیه‌گر فعال باشد

همانطور که نمودار (۱) نشان می‌دهد وقتی که این ویژگی فعال باشد، دقت متوسط برای تمام پروفایل‌های کاربر در روش re-ranking ۰. ۷۶۴۹ است، در حالی که زمانی که غیر فعال است، این دقت به ۰. ۷۲۲۸ کاهش می‌دهد. و در روش پیشنهادی این پژوهش این دقت به مقدار ۰. ۸۱۷۵ رسیده است در جزئیات بیشتر می‌توانید ببینید که در بعضی روزها مانند روزهای ۵ و ۱۱، هنگامی که این سیستم غیر فعال می‌باشد افت ناگهانی دقت مشاهده می‌شود.

در ادامه به مقایسه سه روش مطرح شده در مقاله [۴۲] و روش پیشنهادی در نمودار (۲) برای زمانی‌که لیست علاقمندی‌ها کوچک می‌باشد مانند دیتاست مطرح شده با دوازده موضوع در بیست روز پرداخته خواهد شد این روش‌ها شامل موارد زیر می‌باشد:

Fixed number of interests (Fixed)

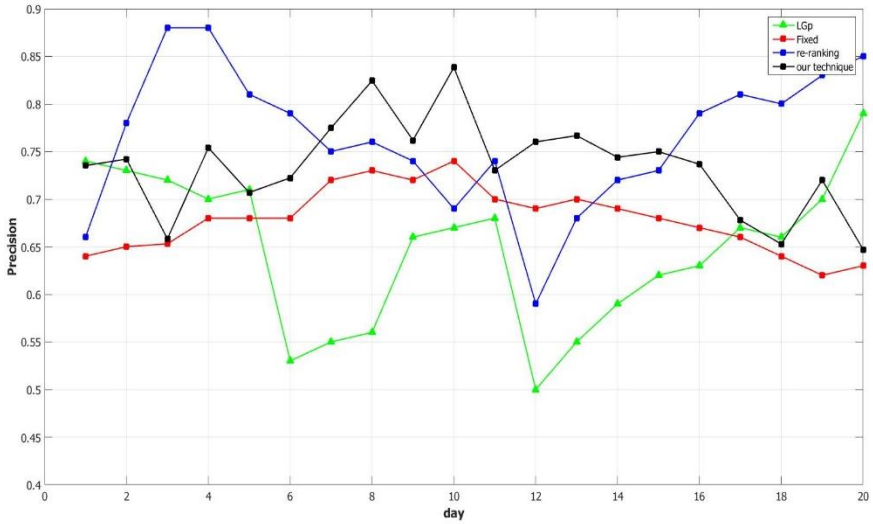
در این روش علاوه بر ذخیره صفحات بازدید شده با زمان زیاد اخرین صفحات بازدید شده در یک بازه زمانی کوتاه نیز ذخیره می‌شود.

Largest gap (LGap)

در این روش ابتدا بزرگترین شکاف را در میان مقادیر فرکانس مرتبط با موضوعات پیدا شده و سپس همه موضوعات با مقادیر فرکانس بزرگتر به عنوان لیست علاقمندی‌های کاربر ذخیره می‌شود.

re-ranking

این روش برای مدل سازی نمایه کاربری پویا با استفاده از میانگین فرکانس و انحراف استاندارد برای کشف علاقه مندی‌های دراز مدت استفاده می‌شود.



نمودار (۲) مقایسه برای چهار الگوریتم

براساس آزمایشات انجام شده بر اساس داده‌های ورودی بر اساس بیست روز و دوازده موضوع به بررسی روش پیشنهادی در مقایسه با الگوریتم‌های ارائه شده در مقاله [۴۲] پرداخته شد که دقت الگوریتم پیشنهادی در مقایسه با الگوریتم‌های دیگر بالاتر می‌باشد.

بحث نتیجه گیری

در این مقاله، یک رویکرد مدل سازی پروفایل کاربری پویا برای شخصی سازی جستجوی صفحات وب براساس علاقمندی ارائه شده است. نقطه آغاز این پژوهش شناسایی تعدادی از مسائل در رویکردهای موجود بود، یعنی محدودیت‌های چنین سیستم‌هایی در برخورد با تغییر علاقمندی‌های کاربر در طول زمان، عدم تعریف روشن از جنبه‌های کوتاه مدت و بلند مدت علاقمندی‌های کاربران و نقاط ضعف در مدل سازی جنبه‌های پویا رفتار کاربران می‌باشد.

این مقاله با هدف رسیدگی به این مسائل به ارائه راهکاری پرداخت. اولاً روش‌های توسعه یافته قادر به مقابله با تغییرات ثابت در علاقمندی‌های کاربران از جمله تغییر ناگهانی این علاقمندی نیستند. ثانیاً، ما روشی طراحی کرده‌ایم که بتواند تا اندازه ای رضایت بخش از علاقمندی‌های کوتاه مدت و بلندمدت کاربر و گذار از یک علاقمندی به علاقمندی دیگر در موضوعات مختلف را به خوبی مدیریت کند و پیشنهادات با علاقمندی بالا به کاربران دهد.

این روش همچنین می‌تواند براساس الگوهای رفتاری کاربران سازگاری داشته باشد روش ارائه شده دارای کاربردهای گسترده‌ای می‌باشد و می‌تواند در تعدادی از حوزه‌های جستجو در اینترنت مورد استفاده قرار گیرد برنامه‌های کاربردی که در آن‌ها امکان پیگیری رفتار کاربر وجود دارد و اطلاعات مربوط به تنظیمات و منافع کاربر می‌تواند استخراج و مدل سازی شود. و

در نهایت، روش ارائه شده با روش‌های دیگر از جهت دقت مورد بررسی قرار گرفت که نتایجی قابل قبولی در غالب نمودارهای ارائه شده نشان داد.

ارائه راهکارهای آتی

به عنوان کارهای پیشنهادی و آتی می‌توان موارد زیر اشاره کرد:

۱-استفاده از الگوریتم‌های خوشه بندی برای مدل سازی رفتار کاربران

۲- افزایش کارایی الگوریتم ارائه شده در ترکیب با الگوریتم پیش بینی مانند شبکه عصبی

منابع

[۱] رعنا فرصتی، محمدرضا میبدی "الگوریتمی مبتنی بر ساختار پیوندی صفحات و اطلاعات استفاده کاربران برای پیشنهاد صفحات وب" دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه صنعتی امیرکبیر، ۱۳۹۴.

[۲] سارا مظعی، "داده کاوی ساختار وب با استفاده از اتوماتای یادگیری توزیع شده و سلولی و کاربردهای آن"، تحت راهنمایی دکتر محمد رضل میبدی، دانشگاه صنعتی امیر کبیر، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، ۱۳۸۷.

[3] A. Shepitsen, et al. Personalized recommendation in social tagging systems using hierarchical clustering. in Proceedings of the 2008 ACM conference on Recommender systems RecSys. 2008. New York, NY, USA.

[4] Adomavicius, G. and A. Tuzhilin, Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. Knowledge and Data Engineering, IEEE Transactions on, 2005. 17(6): p. 734-749.

[5] Ahmad Hawalah "Dynamic user profiles for web personalization" Expert Systems with Applications 42 (2015) 2547–2569

[6] Alan L. Montgomery, Shibo Li, Kannan Srinivasan, and John C. Liechty, " Modeling Online Browsing and Path Analysis Using Clickstream Data" by Alan L. Montgomery, Shibo Li, Kannan Srinivasan, and John C. Liechty, All rights reserved2/24/2015

[7] Asanov, D. , Algorithms and Methods in Recommender Systems. Berlin Institute of Technology, Berlin, Germany, 2011.

[8] B. Sarwar, et al. Item-based collaborative filtering recommendation algorithms. in Proceedings of the 10th international conference on World Wide Web. 2001.

[9] Bindu Madhuri. Ch1, Dr. Anand Chandulal. J2, Ramya. K3 and Phanidra. M4,]" Analysis of Users' Web Navigation Behavior using GRPA with Variable Length Markov Chains" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol. 1, No. 2, March 2011

[10] Bobadilla, J. , et al. , Recommender systems survey. Knowledge-Based Systems, 2013. 46(0): p. 109-132.

[11] Breese, J. S. , D. Heckerman, and C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, in Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. 1998. p. 43-52.

[12] Burke, R. , Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction, 2002. 12(4): p. 331-370.

[13] C. Haruechaiyasak, et al. A dynamic framework for maintaining customer profiles in e-commerce recommender systems. in Proceedings of the 2005 IEEE international conference on e-technology, e-commerce and e-service(EEE'05). 2005. Washington, DC, USA: IEEE: Computer Society.

[14] D. Billsus and M. J. Pazzani, User modeling for adaptive news access. User modeling and user-adapted interaction, 2000. 10: p. 147-180.

[15] D. Goldberg, et al. , Using collaborative filtering to weave an information tapestry. Communications of the ACM, 1992. 35: p. 61-70.

[16] D. Jannach, et al. , Recommender Systems: An Introduction. Cambridge University Press, 2010.
[17] D. Zandi, P. Moradi, and F. Akhlaghian. Evolutionary based matrix factorization method for collaborative filtering systems. in 21st Iranian Conference on Electrical Engineering (ICEE). 2013. Iran, Mashhad.
[18] E. Bojnordi and P. Moradi. A Novel Collaborative Filtering Model based on Combination of Correlation Method with Matrix Completion Technique. in The 16th International Symposium on Artificial Intelligence and Signal Processing (AISP 2012). 2012.
[19] G. Uchyigit and K. Clark, Hierarchical agglomerative clustering for agent-based dynamic collaborative filtering. IDEAL, 2004: p. 827-832.
[20] Goldberg, K. , et al. , Eigentaste: A Constant Time Collaborative Filtering Algorithm. Inf. Retr. , 2001. 4(2): p. 133-151.
[21] Herlocker, J. L. , J. A. Konstan, and J. Riedl, Explaining collaborative filtering recommendations, in Proceedings of the 2000 ACM conference on Computer supported cooperative work. 2000, ACM. p. 241-250.
[22] J. A. Konstan, et al. , GroupLens: applying collaborative filtering to Usenet news. Communications of the ACM, 1997. 40: p. 77-87.
[23] J. Kelleher and D. Bridge. Rectree centroid: An accurate, scalable collaborative recommender. in Proceedings of the fourteenth Irish conference on artificial intelligence and cognitive science. 2003. Citeseer.
[24] J. Wang, A. P. de Vries, and M. J. T. Reinders. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2006. Seattle, Washington, USA.
[25] Jianhan Zhu, Jun Hong, and John G. Hughes" Using Markov Chains for Link Prediction in Adaptive Web Sites" Soft-Ware 2014, LNCS 2311, pp. 60–73, 2014
[26] Jinming, H. Application and research of collaborative filtering in e-commerce recommendation system. in Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference. 2010.
[27] Linden, G. , B. Smith, and J. York, Amazon. com recommendations: item-to-item collaborative filtering. Internet Computing, IEEE, 2003. 7(1): p. 76-80.
[28] M. Pazzani and D. Billsus, Learning and revising user profiles: The identification of interesting web sites. Machine learning, 1997. 27: p. 313-331.
[29] Massa, P. and P. Avesani, Trust-aware recommender systems, in Proceedings of the 2007 ACM conference on Recommender systems. 2007. p. 17-24.
[30] Ortega, F. , et al. , Improving collaborative filtering-based recommender systems results using Pareto dominance. Information Sciences, 2013. 239(0): p. 50-61.
[31] P. Massa and B. Bhattacharjee, Using trust in recommender systems: an experimental analysis. Trust Management, ed: Springer, 2004: p. 221-235.
[32] P. Melville and V. Sindhwani, Recommender systems. Encyclopedia of Machine Learning, 2010: p. 829-838.

[33] P. Melville, R. J. Mooney, and R. Nagarajan. Content-Boosted Collaborative Filtering for Improved Recommendations. in Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI-02). 2002.
[34] P. Resnick, et al. GroupLens: an open architecture for collaborative filtering of netnews. in Proceedings of the 1994 ACM conference on Computer supported cooperative work. 1994.
[35] R. Burke, "Hybrid Recommender Systems: Survey and Experiments," User Modeling and User-Adapted Interaction, vol. 12, pp. 331-370, 2002.
[36] R. J. Mooney and L. Roy. Content-based book recommending using learning for text categorization. in Proceedings of the fifth ACM conference on Digital libraries. 2000.
[37] Resnick, P. , et al. , GroupLens: An open architecture for collaborative filtering of netnews. In Proceedings of the ACM Conference on Computer Supported Cooperative Work CSCW'94, 1994: p. 175-186.
[38] Ricci, F. , et al. , Recommender Systems Handbook. 2010: Springer.
[39] Robert West, Jure Leskovec, " Human Wayfinding in Information Networks"Lyon, France. ACM 978-1-4503-1229- April 16–20, 2012
[40] S. Puntheeranurak and H. Tsuji. A multi-clustering hybrid recommender system. in Proceedings of the 7th IEEE International Conference on Computer and Information Technology. 2007. Washington, DC, USA: IEEE Computer Society.
[41] Sanchez, J. L. , et al. Choice of metrics used in collaborative filtering and their impact on recommender systems. in Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on. 2008.
[42] Sarwar, B. , et al. , Item-based collaborative filtering recommendation algorithms, in Proceedings of the 10th international conference on World Wide Web. 2001, ACM: Hong Kong, Hong Kong. p. 285-295.
[43] Sarwar, B. M. , et al. , Application of dimensionality reduction in recommender systems - a case study, in Proceedings of the ACM WebKDD Workshop. 2000.
[44] Schafer, J. B. , J. Konstan, and J. Riedl, Recommender systems in e-commerce, in Proceedings of the 1st ACM conference on Electronic commerce. 1999, ACM. p. 158-166.