فصلنامه علمی- پژوهشی
فناوری اطلاعات و ارتباطات ایران

سال نهم، شماره‌های ۳۱ و ۳۲، بهار و تابستان ۱۳۹۶
صص: ۵۶- ۴۵

بهبود سیستم‌های توصیه‌گر با کمک وب معنایی

[*]راحله بهشتی‌نژاد [**]محمدابراهیم سمیع [***]علی حمزه
* کارشناسی ارشد مهندسی فناوری اطلاعات، دانشگاه شیراز
** هیات علمی دانشگاه جهرم، گروه مهندسی فناوری اطلاعات
*** هیات علمی دانشگاه شیراز، دانشکده مهندسی برق و کامپیوتر

تاریخ دریافت: ۱۳۹۳/۰۳/۲۴ تاریخ پذیرش: ۱۳۹۳/۱۲/۰۹

چکیده

بشر در زندگی خود به منظور تامین مایحتاج زندگی، همواره از مشاوره و پیشنهادهای دیگران که به‌صورت شفاهی و یا نوشتاری ارائه می‌شوند، بهره گرفته و آن‌ها را در تصمیم‌گیری‌های خود لحاظ می‌نماید. امروزه با پیشرفت فنّاوری و گسترش کسب و کار الکترونیکی در بستر وب‌سایت‌های اینترنتی، فصل جدیدی از زندگی دیجیتال به کمک سیستم‌های توصیه‌گر آغاز گردیده است. مهم‌ترین هدف در این سیستم‌ها، جذب مشتریان و جلب اعتماد آن‌ها از طریق ارائه بهترین و مناسب‌ترین پیشنهاد خرید محصولات، با توجه به علایق و سلایق آن‌ها در میان انبوهی از انتخابات‌ها می‌باشد. در این پژوهش سعی گردیده است، به کمک ارتباطات موجود در هستان‌شناسی DBpedia، اطلاعاتی در ارتباط با حوزه فیلم استخراج گردد. سپس ساختار سیستم توصیه‌گر طراحی و پیاده‌سازی شده و به کمک اطلاعات موجود برروی پایگاه داده MovieLens، عملکرد سیستم توصیه‌گر مورد ارزیابی قرار گرفته است. بنابر ارزیابی‌های انجام شده، مدل پیشنهادی در میان سایر روش‌هایی که به نحوی از وب معنایی بهره می‌برند، از کارایی بالاتری برخوردار است.

واژه‌های کلیدی: سیستم های توصیه‌گر[1]، وب معنایی[2]، هستان‌شناسی[3]، DBpedia.

[1] Recommender System
[2] Semantic Web
[3] Ontology

نویسندهٔ عهده‌دار مکاتبات: محمدابراهیم سمیع samie@shirazu.ac.ir

مقدمه

پیشرفت سریع فناوری اطلاعات و ارتباطات، شبکه ارتباطی جهانی را با افزایش حجم اسناد دیجیتال روبرو کرده است. در این عصر، به دلیل سادگی امکان انتشار اطلاعات در وب و دسترسی بیش از دو سوم از مردم جهان به اینترنت، انسان‌ها به استفاده، تولید و نشر بیش از پیش اطلاعات می‌پردازند [۱]. این افزایش منابع اطلاعاتی، تعداد انتخاب‌های ممکن هر فرد را برای یافتن منابع مورد نیازش افزایش می‌دهد. همچنین به دلیل وجود انتخاب‌های مختلف با کیفیت‌های متفاوت برای هر مورد خاص، سازمان‌دهی نامرتب اطلاعات موجود در وب، تغییرات سریع این اطلاعات و کمبود زمان لازم برای بررسی این منابع، مخصوصاً از نظر میزان صحت اطلاعات هر منبع، تصمیم‌گیری افراد برای گزینش بهترین مرجع با مشکلات بسیاری همراه شده است. در این حالت، حجم اطلاعات در دسترس کاربران وب به حدی زیاد است که قابلیت تصمیم‌گیری و یا به‌روزرسانی اطلاعات، راجع به یک موضوع خاص از کاربران سلب می‌شود. این مشکل که ناشی از پیشرفت سریع فناوری اطلاعات است را گران‌باری اطلاعات[۴] می‌نامند[۲].

برای حل این مشکل، روش‌ها و ایده‌های متنوعی از جمله موتورهای جستجو، خوراک وب[۵] و وب سایت‌های تطبیقی [۲] پیشنهاد شده‌اند که به کاربران برای یافتن، دسترسی، بهره‌برداری و سازمان‌دهی اطلاعات برخط کمک می‌کنند. راه حل دیگر برای مسئله گران‌باری اطلاعات، روش‌های شخصی‌سازی است. سیستم‌های شخصی‌ساز که از یک سیستم توصیه‌گر[۶] بهره می‌برند، سعی بر تنظیم پویای صفحات وب بر اساس علائق شخصی کاربران[۷] دارند.

در فصل اول این مقاله ابتدا به تعاریف مورد نیاز در این حوزه می‌پردازیم. درفصل دوم، پیشینه تحقیق و پژوهش‌های انجام شده در این زمینه مورد بررسی قرار می‌گیرد. در فصل سوم به تفصیل، روش پیشنهادی و نحوه پیاده‌سازی آن بیان گردیده است. فصل چهارم به ارزیابی مدل پیشنهادی پرداخته و نتایج آن در مقایسه با آخرین پژوهش صورت گرفته در این زمینه بیان گردیده است. فصل پنجم به نتیجه‌گیری و فصل ششم به فعالیت‌های آتی پیش رو پرداخته است.

۱- تعاریف

۲- سیستم‌های توصیه‌گر

در راستای پاسخ به نیاز سیستم‌های شخصی‌ساز و خصوصاً با رشد و همه‌گیر شدن آن‌ها، سیستم‌های توصیه‌گر مطرح شدند. این سیستم‌ها را می‌توان فنّاوری شخصی‌سازی شده، برای فیلتر کردن اطلاعات دانست.

در سیستم‌های توصیه‌گر تلاش بر این است تا با حدس زدن شیوه تفکر کاربر (به کمک اطلاعاتی که از نحوه‌ی رفتار کاربر یا کاربران مشابه وی، نظرات آن‌ها و اطلاعاتی که از اقلام[۸] متفاوت وجود دارد)، مناسب‌ترین و نزدیک‌ترین کالا به سلیقه او شناسایی و پیشنهاد گردد.

سیستم‌های توصیه‌گر برای ارائه توصیه‌های خود نیازمند به سه جزء اصلی هستند:

داده‌های زمینه: اطلاعاتی که سیستم پیش از شروع فرآیند توصیه در اختیار دارد.

داده‌های ورودی: اطلاعاتی که در مورد کاربر در حین فرآیند توصیه به سیستم وارد می‌شود.

الگوریتم توصیه: فرآیندی که با کمک داده‌های زمینه و ورودی، به کاربر توصیه می‌دهد. الگوریتم‌های مختلف توصیه، نیازمند داده‌های زمینه و ورودی متفاوتی برای ارائه توصیه هستند [۳].

الگوریتم سیستم های توصیه گر به سه دستهٔ اصلی تقسیم می‌شوند [۴]:

[4] Information Overload
[5] Web Feed
[6] Recommender System
[7] Customizid
[8] Items

مبتنی بر محتوا، فیلترینگ تجمعی (CF) و سیستم های ترکیبی. سیستم های مبتنی بر فیلترینگ نیز به دو دستهٔ کاربر محور و کالا محور تقسیم می‌شوند.

سیستم مبتنی بر محتوا، توصیه‌ها را بر اساس میزان شباهت بین اقلام ارائه می‌کند به‌طوری که شبیه‌ترین اقلام به آن‌هایی را که کاربر قبلاً به آن‌ها رأی[9] مثبت داده و یا آن‌ها را انتخاب کرده است به وی توصیه می‌شود. امروزه این روش به این دلیل که توصیه‌ها را محدود کرده و در معیار انتخاب، فقط ویژگی اقلام را در نظر می‌گیرد، کمتر مورد توجه محققان قرار گرفته است. از سوی دیگر سیستم‌های مبتنی بر فیلترینگ اشتراکی که امروزه خیلی کاربرد دارند، شبیه‌ترین کاربران به کاربر مورد نظر را با روش‌هایی پیدا کرده و بر اساس رأی آن افراد به اقلام مختلف، پرطرفدارترین اقلام را به کاربر توصیه می‌کنند. در سیستم‌های مبتنی بر فیلترینگ اشتراکی، اطلاعات مربوط به رأی هر کاربر به هر قلم جنس و یا خرید یا عدم خرید جنس توسط کاربر وجود دارد؛ و الگوریتم به‌کار گرفته شده باید بر مبنای این اطلاعات بتواند به یک کاربر فعال[10] بر اساس تقاضایی که به سیستم داده است، توصیه‌های دیگری نیز بدهد. علاوه بر این، در بسیاری از سیستم‌ها، از فیلترینگ اشتراکی برای پیشگویی[11] یک مقدار رأی ناشناخته[12]، بر اساس بقیه رأی‌ها استفاده می‌شود.

۱-۲- وب‌معنایی

موج اول پیشرفت وب شامل ارائه حداکثر اطلاعات ممکن به شکلی بود که بتواند به‌صورت مستقیم در قالب زبان نشانه‌گذاری فرامتن[13] برای مخاطب نمایش داده شود. بانک‌های اطلاعاتی هر روز بیش از پیش تلاش می‌کنند تا اطلاعات را به شکلی تولید نمایند که قبل از نمایش برای کاربر، توسط دیگر رایانه‌ها قابل خواندن و پردازش باشد.

آن‌ها برای برچسب‌گذاری داده‌ها، به‌جای زبان نشانه‌گذاری فرامتن از زبان نشانه‌گذاری توسعه‌پذیر[14] بهره می‌برند و برخی از آن‌ها به‌منظور استانداردسازی قالب محتوا، به نحوی که صرف‌نظر از کاربرد نهایی، توسط رایانه‌ها نیز قابل خواندن باشد به استفاده از چارچوب توصیف سند[15] روی آورده‌اند.

در موج جدید پیشرفت تلاش می‌شود تا هر شیء اطلاعاتی موجود در جهان اطلاعات، به‌واسطه موضوع، محل، پدیدآورنده، تاریخ و دیگر ویژگی‌هایش توصیف شود. این نوع اطلاعات پیش‌تر تنها در یک بانک اطلاعاتی ذخیره می‌شد اما اکنون ممکن است در یک سند ذخیره شود.

این موج تازه پیشرفت را «وب معنایی» می‌نامند. وب معنایی شیوه‌ای است برای ایجاد یک وب که در آن رایانه‌ها می‌توانند از شبکه‌ای از داده‌های منبع استفاده کرده، آن‌ها را تعبیر، تحلیل و پردازش کرده و به کاربر ارائه نمایند. این امر ممکن است از بازیابی اسناد گرفته تا جمع‌بندی عناصر برنامه‌ای مختلف برای خلق یک نرم‌افزار کاربردی را دربر گیرد [۵].

با فراگیر شدن موج وب معنایی، حجم کثیری از اطلاعات به قالب‌های سازگار با وب معنایی تبدیل شده‌اند و بسیاری از اطلاعات جدید تولید شده نیز تنها با این قالب قابل دسترس هستند. برای دسترسی به این حجم عظیم اطلاعات، نیازمند نرم‌افزارها و ابزارهایی هستیم که بتوانند اطلاعات را از این پایگاه داده‌های بزرگ استخراج کرده و مورد پردازش قرار دهند.

سیستم‌های توصیه‌گر نیز از این قضیه مستثنا نبوده و می‌بایست خود را با این موج هماهنگ کنند. در غیر این صورت از اطلاعات پروفایل کاربران و اطلاعات پایگاه داده‌های اقلامی که تحت قالب وب معنایی هستند، بی‌استفاده خواهند ماند و این خود، با افزایش حجم این نوع

[9] Rating
[10] Active User
[11] Prediction
[12] Unknown
[13] HTML
[14] XML
[15] RDF

اطلاعات، باعث کاهش کارایی سیستم‌های توصیه‌گر خواهد شد [۶].

حوزه‌ی مفاهیم موجود در بسیاری از سیستم‌های توصیه‌گر فعلی، محدود به دانش موجود در خود سیستم بوده و از منابع عظیم دانش خارج از سیستم، مانند داده‌هایی که در قالب وب‌معنایی هستند و داده‌های پیوندی[۱۶] استفاده نمی‌شود [۶].

اما این اضطرار علاوه بر همگام شدن با تغییرات پایگاه داده‌ها، فوایدی نیز برای سیستم‌های توصیه‌گر دارد. ارتباط معنایی میان هستان‌ها[۱۷] در وب‌معنایی، موجب افزایش ارتباطات معنادار میان علایق کاربران و اقلام مورد نیاز آن‌ها، افزایش ارتباطات معنادار میان کاربران با سلایق مشابه، میان اقلام با ویژگی‌های مشابه و... می‌شود که این موضوع باعث پاسخگویی سیستم‌های توصیه‌گر با سطح بالاتری از هوشمندی می‌گردد.

پیشینه پژوهش

از اوایل دهه نود بحث سیستم های توصیه‌گر و ویژگی‌های آن‌ها مطرح گردید. امروزه سعی در ساخت سیستمی توصیه‌گر با درصد خطای کم و سرعت بالا در تمام شرایط به یکی از پرطرفدارترین حوزه‌های تحقیقاتی دانشگاهی تبدیل شده است.

اما استفاده از وب معنایی به‌عنوان پایگاه دانش برای سیستم‌های توصیه‌گر یک ایده کاملاً ابتکاری و جدید است. روش‌های بسیاری برای افزایش سرعت و کارایی و کاهش خطا در مقابله با مسائل شناخته شده از سیستم‌های توصیه‌گر پیشنهاد شده است که همه آن‌ها بر اساس مبتنی بر محتوا [۷] و فیلترینگ اشتراکی [۸] یا روش‌های ترکیبی می‌باشند.

از طرفی دیگر تعداد زیادی روش، برای مقابله با مسائل شناخته شده در سیستم‌های توصیه‌گر پیشنهاد شده است اما تعداد کمی از آن‌ها وجود دارند که از مقدار عظیم اطلاعات کدگذاری شده در داده‌های وب معنایی بهره‌برداری می‌کنند[۶].

یکی از روش‌هایی که از اطلاعات عظیم در وب معنایی استفاده کرده است، روش فاصله معنایی داده‌های پیوند شده (LDSD) است. این روش که در سال ۲۰۱۰ توسط پاسنت ارائه گردید، از DBpedia به‌عنوان منبع اطلاعات برای محاسبه توصیه‌ها استفاده کرده است.

در مقاله وی زمینه‌های نظری و همچنین پیاده‌سازی یک سیستم توصیه‌گر تحت عنوان dbrec که یک سیستم توصیه‌گر موسیقی بر مبنای DBpedia است شرح داده شده است که توانایی توصیه نمودن ۳۹۰۰۰ هنرمند یکتا را دارد. انگیزه این مقاله این است که بتواند معیار فاصله معنایی را روی منابع منتشر شده روی وب به‌عنوان داده‌های پیوند شده اعمال نماید؛ بنابراین برای رسیدن به این هدف روش فاصله معنایی داده‌های پیوند شده (LDSD) را تعریف می‌کند تا بتواند فاصله دو منبع منتشر شده در داده‌های پیوند شده را که در فاصله [۰،۱] نرمال شده‌اند، محاسبه نماید. در این مقاله تنها فاصله معنایی منابعی محاسبه می‌شود که به طور مستقیم باهم پیوند داده شده‌اند یا حداکثر توسط منبع سومی باهم پیوند داده شده باشند.

نکته قابل توجه در روش ذکر شده توسط پاسنت این است که گسترش معنایی داده‌ها در این روش مورد استفاده قرار نگرفته است. در حالی که گسترش معنایی مفاهیم، کیفیت کلی نتایج را بهبود می‌بخشد. همچنین در این روش تنها بر روی منابع اطلاعاتی موسیقی کار شده است [۹].

میلر و همکارانش نیز در سال ۲۰۰۸ از محتوای متن و ساختار پیوند صفحات ویکی‌پدیا استفاده کرده‌اند تا به شناسایی شباهت بین فیلم‌ها برای وب سایت توصیه‌گر فیلم Netflix Prize بپردازند. با توجه به ساختار بسیار ناقصی که مقالات ویکی‌پدیا دارند، استخراج اطلاعات مفید از آن بسیار دشوار است. این رویکرد تنها بر اساس متن بدون ساختار و ابر پیوند است و نتوانسته دقت سیستم را بهبود بخشد [۱۰].

هیتمن و هایس نیز در سال ۲۰۱۰ استفاده از داده‌های وب معنایی را برای کاهش مسائل به خوبی شناخته شده از سیستم‌های توصیه‌گر پیشنهاد کرده‌اند. از جمله مشکل

[16] Linked Data
[17] Ontology

بزرگ جمع‌آوری داده[18] برای سیستم‌های توصیه‌گر مانند کاربر جدید، قلم جدید و مشکلات تنکی داده[19]. مشکل کاربر جدید زمانی رخ می‌دهد که اطلاعاتی در مورد کاربر جدید وارد شده به سیستم نداریم. مشکل قلم جدید نیز زمانی رخ می‌دهد که اطلاعاتی در مورد قلم جدید وارد شده به سیستم نداریم. ترکیب هردو مشکل کاربر جدید و قلم جدید نیز با عنوان شروع سرد[20] شناخته می‌شود. اگر تعداد رتبه‌هایی که کاربران سیستم به اقلام داده‌اند از تعداد اقلام بسیار کمتر باشد با مشکل تنکی داده‌ها مواجه هستیم. در این روش، آن‌ها شرح داده‌اند که چگونه می‌توان انبوهی از داده‌های شیءگرا را از منابع مختلف جمع‌آوری کنیم، آن‌ها را پردازش کنیم و در سیستم‌های توصیه‌گر تجمعی مورد استفاده قرار دهیم. هیتمن و هایس برای ارزیابی روش پیشنهادی‌شان، داده‌های خود را از یک سیستم توصیه‌گر موسیقی مبتنی بر فیلترینگ جمعی بسته جمع‌آوری نموده‌اند. سپس با کمک داده‌های پیوند شده باز[21] موفق شده‌اند، میزان کارایی و دقت سیستم اولیه را بهبود بخشند. اما وجود ناسازگاری در منابع باز مختلف باعث بالا رفتن هزینه استفاده از آن‌ها می‌شود و ماهیت استفاده از این روش را به چالش می‌کشد.

تعدادی از مطالعات نیز تنها به بررسی نظری فواید استفاده از هستان‌شناسی‌ها در سیستم‌های توصیه‌گر پرداخته‌اند. در این میان لوکا باریانو و همکارانش در سال ۲۰۰۶ نتایج تحقیقاتشان را در مورد نقش هستان‌شناسی‌ها در سیستم‌های توصیه‌گر زمینه آگاه[22] و سیار منتشر کردند. آن‌ها اثبات کرده‌اند که انطباق هستان‌شناسی‌ها برای مدل کردن دامنه اطلاعات، یک قسمت ضروری برای طراحی سیستم‌های توصیه‌گر زمینه آگاه در آینده است و می‌تواند موجب ارائه توصیه‌های بهتری به کاربران گردد [۱۱].

همچنین در سال ۲۰۱۰ آهو سیگ، بامشاد مباشر و روبین بورک، یک سیستم توصیه‌گر ترکیبی حساس بر زمینه جمعی را پیشنهاد دادند که از دانش معنایی در قالب دامنه هستان‌شناسی بهره می‌برد و در آن پروفایل کاربر، مبتنی بر هستان‌شناسی است. در این مقاله ذکر شده است که چگونه پروفایل کاربر مبتنی بر هستان‌شناسی یاد می‌گیرد، مرتباً به‌روز می‌شود و در سیستم توصیه‌گر جمعی به کار برده می‌شود.

نویسندگان مقاله به‌صورت تجربی بر روی دامنه‌ی کتاب در داده‌های طبقه بندی شده در آمازون، نشان داده‌اند که روش‌های مبتنی بر هستان‌شناسی به طور قابل توجهی دقت و تعداد توصیه‌ها را در مقایسه با سیستم فیلترینگ جمعی استاندارد بهبود می‌بخشد؛ و می‌تواند تا حدودی مشکل شروع سرد را در این نوع سیستم‌های توصیه‌گر مرتفع نماید [۱۲]. در این روش صرف ساخت و استفاده از هستان‌شناسی برای اطلاعات پروفایل کاربران مد نظر قرار گرفته است و اطلاعات کتاب همچنان به صورت یک پایگاه داده ثابت می‌باشد. جدیدترین تحقیقات در این زمینه نیز در سال ۲۰۱۲ توسط توماسو دای نویا انجام گرفته است. وی یک سیستم توصیه‌گر مبتنی بر مدل معرفی کرده است که از ابر داده‌های پیوند شده باز[23] مانند DBpedia و LinkedMDB به عنوان منبع اطلاعات مبتنی بر وب معنایی برای استخراج اطلاعات کاربران و اقلام بهره می‌جوید.

در این روش گراف RDFای که دامنه مورد نظر نویسنده را اغنا می‌کند به بردارهای ویژگی که مناسب برای کار دسته-بندی هستند تبدیل می‌شوند. نتایج این پژوهش حاکی از بهبود نتایج توصیه‌ها نسبت به بسیاری از روش‌های قبلی مانند سیستم‌های مبتنی بر محتوا و فیلترینگ تجمعی است. نویا معتقد است که یکی از چالش‌های مهم پیش روی سیستم‌های توصیه‌گر نیاز آن‌ها به استفاده از داده‌هایی است که درقالب وب معنایی منتشر می‌شوند. در این پژوهش بر

[18] Data acquisition
[19] Sparsity
[20] Cold Start
[21] Liking Open Data (LOD)
[22] Context-Aware
[23] Linked Open Data (LOD)

میزان انطباق سیستم‌های مبتنی بر مدل با منابعی که از طریق وب معنایی استخراج می‌شوند تاکید شده است[۶]. نکته‌ای که در این مقاله کمتر به آن توجه شده است انتخاب ویژگی‌های مناسب برای بهبود پاسخ‌های دریافت شده از سیستم توصیه‌گر می‌باشد، همچنین عدم توجه در استفاده از ویژگی‌های ترکیبی مانع رسیدن به پاسخ مطلوب شده است.

چارچوب طرح پیشنهادی

در این پژوهش ما یک سیستم توصیه‌گر مبتنی بر مدل را معرفی می‌کنیم که به‌جای داده‌های استاندارد زمینه خود، از داده‌های پیوند شده در DBpedia استفاده می‌کند. در این روش ما گراف سه‌تایی‌های RDF[۲۴] استخراج شده از DBpedia را به بردارهای ویژگی تبدیل می‌کنیم تا بتوانیم به‌راحتی از تکنیک‌های یادگیری ماشین و به طور خاص SVM استفاده نماییم.

سپس برای هر قلم[۲۵] (هر فیلم موجود در دامنه) از مجموعه دامنه فیلم و برای هر ویژگی آن (نظیر کارگردان، نویسنده، بازیگران و...)، مجموعه منابعی (به طور مثال تمام بازیگران ستاره یک فیلم یا تمامی نویسندگان آن) که به آن ویژگی پیوند داده شده‌اند را استخراج نماییم؛ بنابراین هر قلم با یک بردار چند بعدی در فضا نشان داده می‌شود که هر بعد آن به یک منبع از منابع موجود مربوط است. سپس برای همه‌ی ویژگی‌ها، هر قلم با یک بردار وزن منحصر به فرد نمایش داده می‌شود که وزن هر بردار به میزان ارتباط میان قلم و منبع، بر اساس ویژگی مدنظر است. وزن‌ها با کمک معیار TF-IDF[۲۶] محاسبه گشته‌اند.

پیاده‌سازی طرح پیشنهادی

در این طرح از مجموعه داده‌های مربوط به DBPedia استفاده گردیده است. اطلاعات مورد نظر را از DBPedia با استفاده از زبان SPARQL[۲۷] می‌توان مورد جستجو قرار داد.

داده‌های پایگاه داده DBPedia به عنوان هستان‌شناسی فیلم در نظر گرفته شده‌اند. این پایگاه داده شامل همه کلاس‌ها، خصیصه‌ها و افراد استخراج شده از ویکی‌پدیا است. به همین دلیل می‌تواند به عنوان یک هستان‌شناسی کامل از مفاهیم فیلم در نظر گرفته شود و برای طرح حاضر بسیار مفید خواهد بود.

زبان برنامه‌نویسی انتخاب شده برای اجرای این طرح زبان Java با پلتفرم Eclipse می‌باشد. به همین دلیل کتابخانه‌ای که برای کار با هستان‌شناسی انتخاب می‌شود باید بر مبنای Java باشد. با بررسی انجام شده کتابخانه‌ای کامل حاوی کلاس‌ها و توابع لازم برای کار با یک هستان‌شناسی به نام Jena انتخاب شده است؛ که این کتابخانه به‌صورت متن باز در اختیار عموم قرار دارد.

برای شروع می‌بایست URL فیلم‌هایی که در پایگاه داده MovieLens در مورد آن‌ها نظر داده شده است، مورد جستجو قرار گیرد. بدین منظور، نام فیلم‌های مشترک در مجموعه داده فیلم پایگاه داده MovieLens و هستان‌شناسی DBPedia مورد جستجو قرار گرفت و آدرس URLهای فیلم‌ها در هستان‌شناسی DBPedia استخراج گردید.

به کمک زبان برنامه نویسی Java و کتابخانه کار با اجزا وب معنایی Jena و زبان SPARQL برنامه‌ای نوشته شد که با برقراری ارتباط با هستان‌شناسی DBPedia برای هر URL فیلم استخراج شده در مرحله قبل، به ترتیب اطلاعات زیر استخراج گردید:

۱. نام کارگردان فیلم؛

۲. نام بازیگران اصلی فیلم؛

۳. نام کشور تولید کننده فیلم.

سپس به کمک اطلاعات به‌دست آمده، مجموعه داده‌های زیر استخراج گردیدند. بدین گونه که در صورتی که فیلم مورد نظر خصیصه‌ای را داشته باشد، برای آن خصیصه مقدار یک و در غیر این صورت مقدار صفر می‌گیرد.

مجموعه داده فیلم بر اساس کشور سازنده فیلم.

مجموعه داده فیلم بر اساس کارگردان فیلم.

مجموعه داده فیلم بر اساس بازیگران اصلی فیلم.

مجموعه داده فیلم بر اساس بازیگران اصلی و کارگردان فیلم.

[24] Resource Description Framework
[25] Item
[26] Term Frequency–Inverse Document Frequency
[27] Simple Protocol and RDF Query Language

سپس زیر برنامه‌ای در محیط برنامه نویسی Matlab جهت محاسبه وزن خصیصه‌های هریک از مجموعه داده‌های مرحله قبل بر اساس سیستم وزن دهی TF-IDF نوشته شد و بر روی هریک از مجموعه داده‌های مرحله قبل اعمال گردید.

روش TF-IDF که از رایج‌ترین روش‌های وزن‌دهی ویژگی‌ها به شمار می‌رود، حاصل ترکیب روش‌های مبتنی بر TF و روش‌های مبتنی بر IDF است. که به صورت زیر محاسبه می‌شود [۱۳].

$$w_{ki} = TFIDF(t_k, d_i) = tf(t_k, d_i) * idf(t_k, d_i)$$

ارزیابی طرح پیشنهادی

جهت ارزیابی روش پیشنهادی نیازمند حجم وسیعی از اطلاعات مربوط به نظرات کاربران مختلف در مورد فیلم‌های مختلف می‌باشیم. وب سایت MovieLens یک وب سایت بسیار قوی و مشهور در زمینه توصیه فیلم است. در این سایت کاربران با نظر دادن به فیلم‌هایی که تاکنون دیده‌اند به کمک سیستم‌های توصیه‌گر مبتنی بر فیلترینگ جمعی، توصیه‌هایی را در مورد فیلم‌هایی که می‌تواند مورد علاقه‌شان باشد، دریافت می‌کنند. این سایت هزاران کاربر دارد و میلیون‌ها نظر آنان را در مورد فیلم‌هایی که دیده‌اند گردآوری نموده است؛ و جهت توسعه دانش موجود در زمینه سیستم‌های توصیه‌گر، این اطلاعات را به رایگان منتشر نموده است.

جهت ارزیابی روش مدنظر پایگاه داده ای از نظرات کاربران در مورد فیلم ها از وب سایت MovieLens دریافت شده که شامل یک میلیون نظر از ۶۰۰۰ کاربر در مورد ۴۰۰۰ فیلم است. در این پایگاه داده هر کاربر حداقل در مورد ۲۰ فیلم نظر داده است.

بیان این نکته حائز اهمیت است که ما رتبه‌های بین یک تا پنج کاربران را به حالت باینری تبدیل نمودیم، بدین‌گونه که رتبه‌های ۱ تا ۳ را برابر ۰ که نشان از عدم علاقه کاربر و ۴ و ۵ را برابر ۱ که نشان از علاقه کاربر است قرار دادیم.

سپس زیربرنامه‌ای در محیط برنامه‌نویسی Matlab جهت تهیه یک مجموعه داده برای هر کاربر نوشته شد که در هر سطر این مجموعه داده، خصیصه‌های فیلم‌هایی که کاربر در مورد آن‌ها نظر داده است، با استفاده از مجموعه داده‌های ساخته شده در مرحله پیاده‌سازی و همچنین نظر کاربر در مورد هر فیلم به عنوان برچسب آن فیلم در ستون آخر قرار دارد. بدین ترتیب برای هر کاربر و به کمک مجموعه داده‌های ایجاد شده در مرحله پیاده‌سازی یک سری مجموعه داده تولید شده است که در آن فیلم‌هایی که کاربر دیده است به همراه ویژگی‌های آن فیلم و نظر کاربر درمورد آن فیلم ذخیره گردیده است.

در ادامه در محیط برنامه‌نویسی Matlab زیر برنامه‌ای جهت انتخاب ویژگی‌های مناسب[۲۸] برای دسته‌بندی روی مجموعه داده‌های کاربران نوشته شد.

جهت انتخاب ویژگی‌های مناسب برای دسته‌بندی از تابع آماده موجود در Matlab با نام relieff استفاده شد که برای انتخاب ویژگی‌ها از الگوریتم Relieff استفاده می‌کند. این روش از یک راه حل آماری برای انتخاب ویژگی استفاده می‌کند، همچنین یک روش مبتنی بر وزن است که از الگوریتم‌های مبتنی بر نمونه الهام گرفته است.

سپس الگوریتم دسته‌بندی SVM با 10-Fold Cross Validation بر روی مجموعه داده‌های کاربران اجرا گردید علت انتخاب SVM این بود که، یکی از بهترین تکنیک‌های دسته‌بندی مورد استفاده در دسته‌بندی متن است و با مسئله مورد نظر، برای یادگیری پروفایل کاربر به خوبی منطبق است. همچنین با مشکل طبیعت خلوت بردار ویژگی‌ها و ابعاد بالای فضای داده‌های ورودی به خوبی کنار می‌آید. همچنین از دیگر مزایای SVM در دسته‌بندی متون می‌توان به موارد زیر اشاره کرد:

- معمولاً به انتخاب زیاد پارامترها[۲۹] نیاز ندارد؛
- در مقابل فرایادگیری[۳۰] مقاوم است؛

[28] Feature Selection
29 Term Selection
[30] Over-fitting

- نیازی به تلاش ماشین یا انسان برای تنظیمات زیاد پارامتر[۳۱] در یک مجموعه اعتبارسنجی[۳۲] ندارد.

زمانی که حد تصمیم‌گیری خطی نیست، نیاز است تا داده‌ها به فضایی با ابعاد بالاترانتقال یابند. این امکان به کمک تبدیلات ریاضی که با تبدیلات به کمک کرنل[۳۳] شناخته می‌شوند انجام می‌گیرد. برای انجام این کار سه تابع مهم مورد آزمایش قرار گرفت: ۱. خطی[۳۴] ۲. چند جمله‌ای[۳۵] ۳.RBF[۳۶]

و نهایتاً RBF انتخاب شد. چراکه در دامنه پژوهش بهترین نتایج را کسب نمود.

همچنین برای پیاده سازی SVM از توابع آماده Matlab استفاده گردید؛ و از نتایج SVM برای ساخت مدل‌های منطقی که قادر به تخمین صحیح دسته‌ها هستند استفاده شد. خروجی سیستم توصیه‌گر می بایست یک فهرست رتبه‌بندی بین مقادیر ۰ و ۱ باشد که از مدل‌های منطقی به‌دست می‌آمد.

نکته حائز اهمیت اینکه چارچوب روش پیشنهادی منحصر به دسته‌بندی SVM نیست و می‌توان از سایر روش‌های دسته‌بندی نیز استفاده نمود.

همچنین برای اینکه امکان مقایسه‌ای برای استفاده از اطلاعات هستان‌شناسی DBPedia و روش‌های متعارف و معمول نظیر آنچه وب سایت MovieLens از آن بهره می‌جوید وجود داشته باشد، مجموعه داده‌های فیلم این سایت را نیز مورد جستجو قرار داده و پایگاه داده‌ای برای هر کاربر با استفاده از این مجموعه داده‌ها ایجاد گردید و الگوریتم دسته‌بندی SVM بار دیگر نیز بر این پایگاه داده‌ها اعمال گردید، به عبارت دیگر آنچه تا کنون شرح داده شد بر روی داده‌هایی که در مورد فیلم‌ها از پایگاه داده MovieLens به دست آمد اعمال گردید، هدف از انجام این کار، مقایسه نتیجه استفاده از داده‌های پیوند شده باز در هستان‌شناسی‌ها در مقایسه با داده‌های ایستا و محدود می‌باشد.

۵-۱- نتایج

در این بخش، نتایج به‌دست آمده از پیاده‌سازی روش پیشنهادی مورد بررسی قرار می‌گیرد. این بررسی در قالب روش‌های ارزیابی یادگیری ماشین همانگونه که در جدول ۵-۱ آورده شده است انجام گردیده است. همچنین برای ارزیابی چارچوب پیشنهادی از منحنی مشخصه عملکرد سیستم (ROC[۳۷]) (مطابق نمودار ۵-۱) استفاده شده است.

۵-۲- مقایسه طرح پیشنهادی با روش موجود

با مقایسه نتایج حاصله از بخش پیشین، بهینه‌ترین نتیجه از میان حالت‌های مورد بررسی، مربوط به مجموعه داده فیلم بر اساس بازیگران اصلی و کارگردان فیلم می‌باشد. در این بخش، این نتیجه را با نتیجه ارائه شده در یکی از آخرین مقالاتی که در این زمینه منتشر گردیده مقایسه می‌نماییم. نکته قابل توجه این است که در مقاله مذکور نتایج بهتری نیز حاصل گردیده است که علت آن استفاده از سطح دوم ویژگی‌های یک هستان‌شناسی است. لذا مقایسه تنها با مجموعه داده‌ای که مربوط به موضوع فیلم و در سطح اول است انجام پذیرفته است. سپس مقایسه‌ای بین نتایج اعمال چارچوب طرح پیشنهادی بر روی داده‌های استخراج شده از DBPedia و داده‌های استخراج شده از وب سایت MovieLens صورت گرفت که به طور مشخص می‌توان نتایج بهتری در حالتی که از داده‌های DBPedia استفاده گردیده است مشاهده نمود.

[31] Parameter Tuning
[32] Validation Set
[33] kernel trick
[34] Linear
[35] Polynomial
[36] Radial Basis Function
[37] Receiver Operating Characteristic

در جدول ۵-۲، مقایسه شاخص‌های ارزیابی روش پیشنهادی و روش توماسو دای نویا و روشی که چارچوب پیشنهادی را بر روی داده‌های فیلم وب‌سایت MovieLens اعمال نمودیم طبق مجموعه داده فیلم براساس بازیگران اصلی و کارگردان فیلم آورده شده است. همانگونه که مشاهده می‌شود، نتایج حاصله از روش پیشنهادی بصورت قابل ملاحظه‌ای بهینه‌تر از روش توماسودای نویا و همچنین روش توماسو دای نویا نیز بهینه‌تر از حالتی است که از داده‌های فیلم MovieLens استفاده گردید. این موضوع در نمودار ۵-۲ نیز بخوبی نمایش داده شده است.

۵-۱- نتیجه‌گیری

امروزه وب‌داده‌ها شامل حجم عظیمی از اطلاعات ساختاریافته است که قابل استفاده برای کاربران نهایی و ارائه‌دهندگان سرویس‌های مختلف می‌باشد. در این پژوهش نشان داده شد که چگونه دانش کد شده در ابر داده‌های پیوند شده باز موجود در هستان‌شناسی‌ها می‌تواند در نتایج اخذ شده در سیستم توصیه‌گر تأثیرگذار باشد. یکی از فواید استفاده از داده‌های پیوند شده باز برای سیستم‌های توصیه‌گر، کاهش ابعاد مسئله‌ی تحلیل محتوای محدود شده[۳۸] است.

در واقع، عدم تجانس موضوعات و زمینه‌های ارائه شده در ابر داده‌ها و همچنین ماهیت پیوسته آن، موجب انتخاب آسان و بهره‌برداری از ویژگی‌های/خواص جدید و متنوع برای یک دامنه خاص می‌شود. ماهیت هستان‌شناسی داده‌ها که در داده‌های پیوند شده باز وجود دارد اثبات کرده است که می‌تواند دقت کلی سیستم را بهبود بخشد. نتایج ارائه شده، یک گام اولیه از بررسی و تحلیل جامع جنبه‌های متفاوت استفاده از داده‌های پیوند شده باز روی سیستم‌های توصیه‌گر است؛ و نتایج امیدوارکننده این پژوهش، راه را برای بسیاری از پژوهش‌های نوین در زمینه سیستم‌های توصیه‌گر باز خواهد نمود.

۶- آینده پژوهش

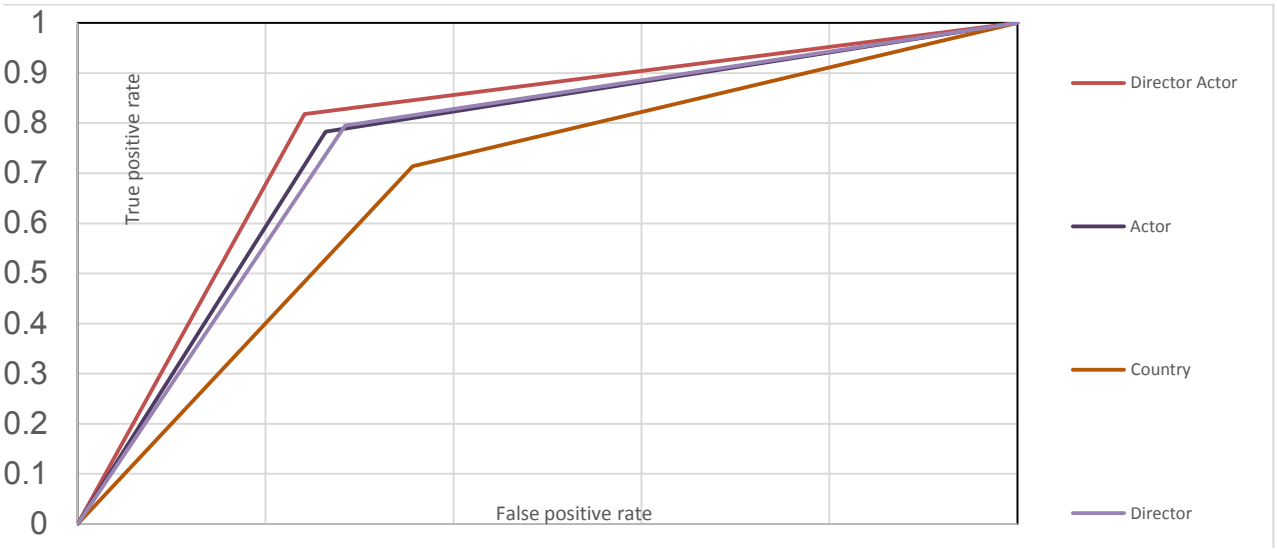
در این پژوهش به سطح اول اطلاعات در هستان‌شناسی تاکید شده است، حال آنکه می‌توان برای نشان دادن میزان قدرت داده‌های پیوند شده باز در هستان‌شناسی‌ها داده‌ها را در سطح دوم نیز مورد بررسی قرار داد، در واقع هدف این است که ارتباطات معنادار میان داده‌ها در هستان‌شناسی با عمق بیشتری مورد کنکاش قرار گیرد تا بتوان با ترکیب ویژگی‌ها و استفاده از این ارتباطات معنادار در سطوح پایین‌تر نتایج بهتری حاصل نمود. به طور مثال برای کارگردان یک فیلم تمام فیلم‌هایی را که کارگردانی کرده است استخراج کرده آن‌ها را خوشه بندی نموده و مجموعه داده فیلم را با آن تشکیل دهیم.

چارچوب این پژوهش مختص فیلم نبوده و می‌توان با انتخاب سایر دامنه‌های متداول نظیر موسیقی، کتاب و ...، علاوه بر اثبات مؤثر بودن روش فوق برای آن دامنه‌ها، به نقاط ضعف و قوت این روش و نحوه بهبود و تقویت آن نیز دست یافت.

[38] Limited Content Analysis

جدول۵-۱: توضیح جدول‌های شاخص‌های ارزیابی

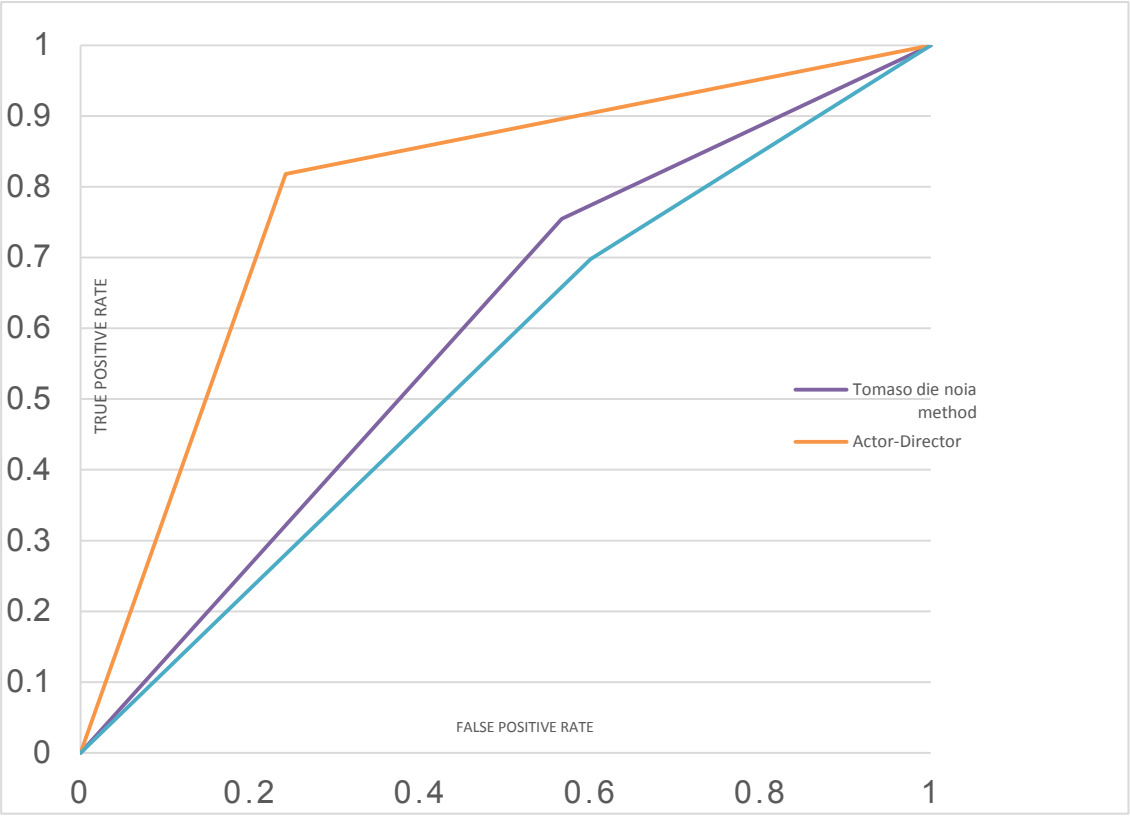
مجموعه داده	AUC	FPR	ACC	PRE	REC	F-Measure
	سطح زیر منحنی مشخصه عملکرد سیستم	نسبت منفی غلط	دقت	درصد	یادآوری	امتیاز F
کشور سازنده فیلم	70.870±7.9	46.053	71.430±9	67.267±15.9	62.791±16.4	**64.154±14.1**
کارگردان فیلم	76.491±5.2	32.203	77.226±5	75.390±5.5	72.115±10.9	**73.471±4.5**
بازیگران اصلی فیلم	76.451±5.8	26.917	78.008±8.7	75.327±6.3	69.712±7.3	**76.915±4.6**
بازیگران اصلی و کارگردان فیلم	78.836±8	24.327	79.437±5.8	80.682±6.6	79.978±10.8	**79.898±4.9**



نمودار ۵-۱: ROC برای مجموعه داده های انتخابی

جدول ۵-۳: مقایسه شاخص‌های ارزیابی روش پیشنهادی با مجموعه داده فیلم براساس بازیگران اصلی و کارگردان فیلم (با استفاده از اطلاعات وب معنایی) و روش توماسو دای نویا و استفاده از چارچوب SVM بر روی داده‌های فیلم سایت MovieLens (بدون استفاده از اطلاعات وب معنایی)

	AUC	FPR	ACC	PRE	REC	F
روش پیشنهادی با استفاده از اطلاعات وب معنایی	78.836±8	24.327	79.437±5.8	80.682±6.6	79.978±10.8	79.898 **±4.9**
روش توماسو دای نویا	59.420±8.4	41.552	61.060±9.3	67.040±9.8	59.431±21.1	67.630 **±9.3**
استفاده از چارچوب SVM بر روی داده‌های فیلم سایت MovieLens (بدون استفاده از اطلاعات وب معنایی)	56.330±5.6	47.592	57.088±8.3	62.080±7.7	57.833±32.3	61.721 **±6.6**



نمودار ۵-۲: منحنی مشخصه عملکرد سیستم برای روش توماسوی‌دای نویا در مقایسه با مجموعه داده فیلم براساس براساس بازیگران اصلی و کارگردان فیلم (با استفاده از اطلاعات وب معنایی) و روش استفاده از چارچوب SVM بر روی داده‌های فیلم سایت MovieLens (بدون استفاده از اطلاعات وب معنایی)

منابع

1. H. Shimazu, "ExpertClerk: navigating shoppers' buying process with the combination of asking and proposing," in Proceedings of the 17th international joint conference on Artificial intelligence-Volume 2, 2001, pp. 1443–1448.

2. M. Perkowitz and O. Etzioni, "Adaptive web sites," Commun. ACM, vol. 43, no. 8, pp. 152–158, 2000.

1.B. Heitmann and C. Hayes, "Using Linked Data to Build Open, Collaborative Recommender Systems," in AAAI Spring Symposium: Linked Data Meets Artificial Intelligence, 2010, pp. 76–81.

۲.ن. . . مقدم چرکری آرش نیک نفس علی اکبر نیک نفس, "سیستم توصیه گر مبتنی بر روش PROMETHEE IIبرایدسته های‌مختلف‌اقلام با تکرار خرید پایین14 "،امین کنفرانس ملی سالانه انجمن کامپیوتر ایران .تهران, ۲۰۰۹.

3.S. E. Middleton, D. De Roure, and N. R. Shadbolt, "Ontology-based recommender systems," in Handbook on Ontologies, Springer, 2009, pp. 779–796.

4. T. Di Noia, R. Mirizzi, V. C. Ostuni, and D. Romito, "Exploiting the web of data in model-based recommender systems," in Proceedings of the sixth ACM conference on Recommender systems, 2012, pp. 253–256.

5. M. Pazzani and D. Billsus, "Content-based recommendation systems," Adapt.

web, 2007.

8. J. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," Proc. Fourteenth …, 1998.

9. A. Passant, "Dbrec—music recommendations using DBpedia," *Semant. Web–ISWC 2010*, 2010.

10. J. Lees-Miller and F. Anderson, "Does Wikipedia Information Help Netflix Predictions?," *Mach. Learn. ...*, 2008.

11.L. Buriano and M. Marchetti, "The role of ontologies in context-aware recommender systems," *... , 2006. MDM 2006. ...*, 2006.

12.A. Sieg, B. Mobasher, and R. Burke, "Ontology-based collaborative recommendation," *Computing*, 2010.

Netflix Predictions?," *Mach. Learn. ...*, 2008.

13. G. Salton and C. Yang, "On the specification of term values in automatic indexing," *J. Doc.*, 1973.

14. I. Witten, E. Frank, and M. Hall, *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. 2011.

15. A. Unler and A. Murat, "A discrete particle swarm optimization method **for** feature selection in binary classification problems," *Eur. J. Oper. Res.*, 2010.

16. P. Hitzler, M. Krotzsch, and S. Rudolph, *Foundations of semantic web technologies*. 2011.