یک سامانۀ پیشنهادگر محتوا-مشارکتی مبتنی بر خوشه‌بندی و هستان‌شناسی

پیام بحرانی[۱]، بهروز مینایی بیدگلی[۲]، حمید پروین*[۳و۴]، میترا میرزارضایی[۱] و احمد کشاورز[۵]

[۱]گروه مهندسی کامپیوتر، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران

[۲]دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، تهران، ایران

[۳]گروه مهندسی کامپیوتر، واحد نورآباد ممسنی، دانشگاه آزاد اسلامی، نورآباد ممسنی فارس، ایران

[۴]باشگاه پژوهشگران جوان و نخبگان، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

[۵]گروه مهندسی برق، دانشکده مهندسی سیستم‌های هوشمند و علوم داده، دانشگاه خلیج فارس، بوشهر، ایران

چکیده

سامانه‌های پیشنهادگر سامانه‌هایی هستند که در گذر زمان یاد می‌گیرند که هر فرد یا مشتری به‌احتمال چه کالا یا قلمی را می‌پسندد و آن را به او پیشنهاد می‌دهند. این سامانه‌ها اغلب بر اساس رفتارهای مشابه از دیگر افراد (به‌احتمال مشابه) عمل می‌کنند. به‌طور کلی یافتن افراد مشابه بسیار زمان‌بر و نادقیق است. به همین دلیل برخی از روش‌ها، به روش‌های ترکیبی رو آورده‌اند. در سامانه پیشنهادگر ترکیبی پیشنهادی، از یک سامانه دو مرحله‌ای استفاده کرده‌ایم که در مرحله نخست، دو مدل پیش‌بینی‌های خود را انجام داده، سپس در مرحله دوم به‌وسیله یک مؤلفه ترکیب‌گر، نتایج دو بخش مرحله نخست با یکدیگر ترکیب شده و نتایج به‌دست‌آمده را به‌عنوان نتایج نهایی سامانه به ما ارائه می‌دهد. در بخش نخست، یک سامانه مبتنی بر پرکردن مقادیر گم‌شده، مقادیر خالی در ماتریس امتیازدهی را پر می‌کند. برای این مهم، از بین روش‌های پرکردن داده‌های گم‌شده، روشی را که با پرکردن مجموعه‌داده در شرایط بسیار تُنُک سازگار بود طراحی کرده و سپس آن را به روش خودمان تعمیم داده‌ایم. در این راستا یک روش مبتنی بر خوشه‌بندی فاصله‌گری ارائه کرده‌ایم. بخش دوم خود یک سامانه پیشنهادگر ترکیبی هستان‌شناسی پایه است. به‌کمک یک روش اندازه‌گیری شباهت ابتکاریِ هستان‌شناسی پایه، مشابهت قلم-قلم‌ها، کاربر-کاربرها، و کاربر-قلم‌ها را اندازه‌گیری می‌کنیم. به‌کمک این ماتریس مشابهت، کاربرها و قلم‌ها را خوشه‌بندی و سپس برای هر کاربر، کاربرها و قلم‌های شبیه به آن را به‌عنوان یک ویژگی جدید در پروفایل کاربر ذخیره می‌کنیم. این کار به ما کمک می‌کند که در آینده، سرعت یافتن کاربرهای مشابه و قلم‌های مشابه را بالا ببریم. درحقیقت بر اساس این ویژگی، سرعت کل کار را افزایش داده‌ایم. از آنجایی که ما هدف خود را ساختن سامانه‌ای که یک موازنه بین دو معیار دقت و سرعت را برقرار کند قرار داده‌ایم، با استفاده از یک مجموعه‌داده واقعی، از این دو معیار جهت ارزیابی سامانه پیشنهادی استفاده می‌کنیم. نتایج مقایسۀ روش پیشنهادی ما با برخی روش‌های مشابه به‌روز ارائه‌شده در این حوزه حاکی از آن است که روش ما از روش‌های سریع، کندتر، اما از آنها دقیق‌تر است؛ برای مثال سریع‌ترین روش مبتنی بر K-means به‌طور متوسط کمتر از ۱ (حدود ۰.۱۱) ثانیه برای اجرا مصرف می‌کند ولی Recall آن حدود ۰.۳۰ است؛ در حالی روش ما حدود ۰.۴ ثانیه زمان لازم دارد ولی Recall ما ۰.۸۰ است. همچنین این نتایج بیان‌گر این موضوع است که روش پیشنهادی از دقیق‌ترین روش‌ها (از جمله OtopN با Recall حدود ۰.۶۵ و زمان مصرفی حدود ۰.۵۸ ثانیه)، سریع‌تر و کیفیت آن نیز قابل رقابت و یا حتی بهتر است.

واژگان کلیدی: سامانه پیشنهادگر، هستان‌شناسی، پالایش حافظه پایه، پالایش مدل پایه، خوشه‌بندی، *k-NN*.

A Content-Collaborative Recommender System based on clustering and ontology

Payam Bahrani[1], Behrouz Minaei-Bidgoli[2], Hamid Parvin[3,4]*,
Mitra Mirzarezaee[1]& Ahmad Keshavarz[5]

[1]Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran

[2]School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran

[3]Department of Computer Engineering, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran

* **Corresponding author**

* نویسندۀ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۴۰۲/۲/۳ • تاریخ پذیرش: ۱۴۰۲/۹/۱۹ • تاریخ انتشار: ۱۴۰۲/۱۰/۲۴ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.20.3.197]

4Young Researchers and Elite Club, Yasooj Branch, Islamic Azad University, Yasooj, Iran
5Department of Electrical Engineering, Faculty of Intelligent Systems Engineering and Data Science, Persian Gulf University, Bushehr, 75169, Iran

Abstract:
Recommender systems are systems that, over time, learn what product(s) or item(s) each person or customer is (are) likely to like and recommend it (them) to him/her. These systems often operate based on similar behaviors from other (possibly similar) people. Finding similar people is generally a highly time-consuming process due to the large number of users and inaccurate due to the lack of information. For this reason, some methods have resorted to increasing speed. On the other hand, some other methods have added additional information so that they can increase the accuracy of finding similar or neighboring users. Some others have resorted to hybrid methods. Recently, by the use of basic clustering methods, which is based on finding the most similar neighbors with the help of users' clustering, as well as by using basic content analysis methods and sometimes adding ontology to these methods, researchers have been able to take the advantage of these methods in order to solve some of the above challenges acceptably. In the proposed hybrid recommender system, we have used a two-stage system in which, in the first stage, two models of predictions are made, then in the second stage, by a combining component, the results of the first two parts are combined and the obtained results are given to us as the final results of the system. In the first part, a system based on imputation of missing values fills in the blanks in the scoring matrix. For this end, among the methods of the missing data imputation, we designed a method that was compatible with filling the data set in very sparse conditions, and then generalized it to our own method. In this regard, we have proposed a method based on the grey distance clustering. In the second part, which itself is a hybrid ontology-based recommender system, we first extract the information of each item with the help of a web crawler, then based on a basic article, we produce our own limited ontology, and after that we apply our proposed method. Then, with the help of a proposed method, we improve the ontology structure, thus increasing the accuracy of measuring semantic similarity between the items and users in later stages, and significantly improving the effectiveness of the created recommendations. It should be noted that this ontology is not comprehensive. Finally, we measure the similarity of item-items, user-users, and user-items using an innovative basic ontology similarity measurement method. By the use of this similarity matrix, we cluster users and items, and then store similar users and items as a new feature in the user/item profile for each user/item. This will help us speed up the process of looking for similar users and similar items in the future. In fact, based on this feature, we have increased the speed of the whole work. Since we have set our goal to build a system that makes a balance between the two criteria of accuracy and speed, we use these two criteria to evaluate the proposed system using a real data set. The results of comparing our proposed method with some up-to-date similar methods presented in this field (using the same data set) implies that our method is slower than fast methods, although it is more accurate than them; for example Recall of the fastest method with 0.11 second per prediction is 0.30 while our method consumed time is 0.40 and its Recall is 0.80. These results also suggest that the proposed method is faster than accurate methods and its quality is more competitive or even better than them; for example the OTopN consumes about 0.58 seconds and has a Recall of 0.65.

Keywords: Recommender System; Ontology; Memory-based Filtering; Model-based Filtering; Clustering; *k-NN.*

۱- مقدمه

پیداکردن اطلاعات در تارنماهای بزرگ یک روند دشوار و وقت‌گیر است. رویکرد استفاده از سامانه‌های پیشنهادگر در خط مقدم بازیابی اطلاعات و سامانه‌های پالایش اطلاعات ظاهر می‌شود. این سیستم‌ها به‌منظور توصیه کالاها/خدمات مورد نیاز کاربران (بدون نیاز به جستجوی صریح) توسعه داده شده‌اند. تاریخچه سامانه‌های پیشنهادگر به سال ۱۹۷۹، در ارتباط با علوم شناختی [1]، برمی‌گردد. این سامانه‌ها در سایر زمینه‌های کاربردی حائز اهمیت هستند. بر اساس نیاز افراد، سامانه‌های پیشنهادگر به آن‌ها در یافتن اقلام مناسب کمک می‌کنند [3, 2]. سامانه‌های پیشنهادگر با ارائه پیشنهادهایی (تولید مجموعه‌ای از پیشنهادهای سفارشی) مطابق با علایق کاربران، به آن‌ها کمک می‌کنند تا اطلاعات مورد نیاز خود را پیدا کنند. سامانه‌های پیشنهادگر در تجارت الکترونیک به‌طور گسترده مورد استفاده قرار می‌گیرند.

در این مقاله قرار است، سامانه پیشنهادگر تلفیقی ارائه شود که از ابزار هستان‌شناسی و شباهت‌سنج معنایی به‌کمک wordnet برای حل مشکل شروع سرد استفاده کند. همچنین مقیاس‌پذیری کار، مورد تحلیل قرار خواهد گرفت. در ضمن از آن‌جایی‌که سامانه‌های پیشنهادگر (محتوا پایه و مشارکتی) در مواجهه با مشکل پیشنهاد اقلام جدید و به‌طورکامل بی‌ربط از سوابق گذشته کاربر، نمی‌توانند به‌صورت مطلوبی عمل کنند، در این مقاله به‌سمت ایجاد

[DOI: 10.61186/jsdp.20.3.197]

شباهت‌های مشارکتی- معنایی اقلام رفته و سعی در بهبود مشکل یادشده می‌شود. همچنین رویکردی برای بهبود پروفایل کاربران به‌کمک سازوکار بالا استفاده خواهد شد. در حال حاضر، سامانه پیشنهادگری که در پردازش ارائه پیشنهادها، تمامی موارد بالا را مدنظر قرار دهد، وجود ندارد؛ ازاین‌رو، این پژوهش سعی در توسعه سامانه پیشنهادگر ترکیبی جدیدی مبتنی بر پالایش محتوا پایه، پالایش مشارکتی و هستان‌شناسی دارد.

در بخش پالایش مشارکتی، از روش خوشه‌بندی، پروفایل کاربر بر اساس هستان‌شناسی، هستان‌شناسی اقلام، شباهت معنایی بین دو هستان‌شناسی و الگوریتم رده‌بندی پیشنهادشده، برای غلبه بر مشکلات پراکندگی و شروع سرد استفاده می‌شود. این ادغام، به‌ترتیب، مقیاس‌پذیری و دقت سامانه را بهبود می‌بخشد. از سوی دیگر، هستان‌شناسی بر مبنای اقلام و شباهت معنایی در پالایش محتوا پایه اعمال می‌شوند. برای بهبود دقت در اندازه‌گیری شباهت معنایی، یک روش ابتکاری در این بخش برای سنجش درجه $IS - A$ بین دو گره هستان‌شناسی اقلام که منجر به ساخت یک فهرست پیشنهادی دقیق‌تر برای کاربر فعال می‌شود، مورد استفاده قرار می‌گیرد.

به‌طور کلی جنبه‌های نوآوری در روش پیشنهادی این پژوهش عبارتند از:

- ارائه یک مدل ارزیابی شباهت معنایی بین-اقلامی،
- بهبود سامانه پیشنهادگر با استفاده از هستان‌شناسی،
- بهبود سامانه پیشنهادگر در مواجهه با پدیده شروع سرد،
- بهبود زمان اجرای سامانه پیشنهادگر،
- استفاده از روش‌های جای‌گذاری مقادیر گمشده مناسب برای بهبود کارایی سامانه‌های پیشنهادگر.

بخش‌بندی مقاله حاضر بدین صورت است که: در بخش ۱ این مقاله، مقدمه و توضیحات مختصری راجع به اهداف و نوآوری مقاله آورده شده است. در ادامه در بخش ۲، ادبیات پژوهش و کارهای مرتبط ارائه و در بخش ۳ روش پیشنهادی ارائه شده است. بخش ۴ شامل آزمایش‌های مربوط به روش پیشنهادی و نتایج تجربی و درنهایت در بخش ۵، نتیجه‌گیری کلی و کارهای آینده مورد بحث واقع شده است.

۲- ادبیات و پیشینهٔ پژوهش

در این بخش به بیان و بررسی نتایج روش‌هایی پرداخته می‌شود که ارتباط بسیار نزدیکی به روش پیشنهادیِ ارائه‌شده در این مقاله دارند و درنهایت در بخش چهارم، برخی از آنها را مورد مطالعه تجربی قرار می‌دهیم.

سامانه‌های پیشنهادگر درکل به سه دسته تقسیم می‌شوند: پالایش مشارکتی[1]، پالایش محتوا پایه[2] و روش ترکیبی[3]. تکنیک پالایش مشارکتی دارای محبوبیت بیشتری نسبت به روش‌های دیگر است. از این روش در بسیاری از تارنماهای خرید برخط استفاده می‌شود [6-4]. رمز موفقیت سامانه‌های پیشنهادگر پالایش مشارکتی پایه، در توانایی ایجاد ارتباط معنادار بین افراد و محصولات مورد علاقه آنها به‌منظور کمک به کاربر فعال در خریدهای آینده است. در این روش از شباهت بین تجربیات گذشته (پروفایل) یک کاربر خاص و علایق کاربران دیگر، برای تشخیص کاربران مشابه (همفکر) با کاربر فعال استفاده خواهد شد. سپس به کمک اطلاعات به‌دست آمده، سامانه پیشنهادگر اقدام به ارائه پیشنهادات یا پیش‌بینی‌هایی در خصوص کاربر فعال می‌کند.

در روش پالایش مشارکتی [7]، کاربر فعال فهرستی از اقلامی که سایر کاربران با سلیقه‌های مشابه، در گذشته دوست داشتند، به‌عنوان پیشنهادِ سامانه، دریافت خواهد کرد. تمامی روش‌های پالایش مشارکتی پایه نیاز به رتبه‌بندی گذشته از طرف کاربران، به‌منظور پیش‌بینی و مطابقت پیشنهاد اقلام در آینده برای کاربر فعال دارند [8]. برای انجام این کار، می‌بایست شباهت بین کاربران و اقلام، با استفاده از روابطی که قادر به اندازه‌گیری فاصله (شباهت) میان آنها هستند، محاسبه شود. از آنجا که پالایش مشارکتی به دو روش کاربر پایه و قلم پایه طبقه‌بندی می‌شود، بر این اساس محاسبه شباهت در هر کدام از روش‌های یادشده، متفاوت خواهد بود [9].

در روش پالایش محتوا پایه، مشخصات اقلام مورد توجه قرار می‌گیرند. این روش متکی به پروفایل کاربر و ویژگی‌های اقلام موجود در سامانه است. در این روش، به‌منظور ساخت مدل پیش‌بینی علایق کاربران، از ویژگی‌های محتوا پایه‌ی اقلام استفاده می‌شود. همچنین پروفایل هر کاربر، بر اساس رتبه‌بندی داده شده توسط آن کاربر شکل می‌گیرد [10]. پالایش محتوا پایه، از اقلام مشاهده‌شده در پروفایل کاربر یاد می‌گیرد و بر اساس آن، پیشنهادهایی به کاربر ارائه می‌کند.

علاوه‌بر روش‌های بالا، روش‌های ترکیبی نیز برای توسعه الگوریتم‌های سامانه‌های پیشنهادگر [11] ارائه شده است.

[1] Collaborative Filtering
[2] Content-Based Filtering
[3] Hybrid

[DOI: 10.61186/jsdp.20.3.197]

به‌طور کلی، روش‌های محتواپایه به سه دسته عمده تقسیم می‌شوند: روش دسته‌بندی[4]، رویکرد فضای برداری[5] و رویکرد هستان‌شناسی پایه[6]. در رویکرد فضای برداری، از بردارهای ویژگی وزن‌دار، که بر اساس ویژگی‌های اقلام موجود در پروفایل کاربر به وجود آمده‌اند جهت یافتن موارد مشابه و ارائه پیشنهادات در آینده به کاربر فعال، استفاده می‌شود. به‌عنوان مثال یکی از شناخته شده‌ترین روش‌های موجود که بر اساس رویکرد مذکور عمل می‌کند، روش Rocchio است [12].

پالایش مشارکتی یکی از روش‌های به‌کاررفته در سامانه‌های پیشنهادگر، با استفاده از دو روش مختلف حافظه‌پایه و مدل‌پایه، است [13,14]. در روش حافظه‌پایه، نیاز به کل رتبه‌بندی‌های داده‌شده به اقلام است. تمامی رتبه‌بندی‌ها در ماتریسی به نام ماتریس قلم-کاربر به‌منظور یافتن همسایگانی از کاربر فعال، جهت ارائه پیشنهادهایی متناسب با علایق کاربر، ذخیره می‌شود. در مقابل، روش مدل پایه متکی بر روش‌های یادگیری ماشین است، و در حقیقت با استفاده از رتبه‌بندی‌های ذخیره‌شده در ماتریس قلم-کاربر، اقدام به ساخت یک مدل پیش‌بینی رتبه اقلام جهت ارائه پیشنهادها به کاربر فعال، می‌کند. برخی از مهمترین روش‌های شناخته شده یادگیری ماشین برای این رویکرد، عبارتند از خوشه‌بندی [15]، pLSA [16]، روش SVD [17] و یادگیری ماشین بر اساس گراف [18].

از آنجایی که تکنیک‌های حافظه پایه بسیار ساده و قابل فهم هستند و همچنین پیاده‌سازی راحتی دارند و از طرفی در دنیای تجارت الکترونیک بازده خوبی از خود نشان داده‌اند، می‌توان گفت یکی از بهترین گزینه‌ها جهت به‌کارگیری در سامانه‌های پیشنهادگر هستند. با تمام این تفاصیل بایستی اذعان کرد که این روش اغلب در برنامه‌های کاربردی بزرگ ناموفق عمل می‌کند.

یکی از مشکلاتی که این روش با آن روبه‌رو است مسئله پراکندگی داده‌ها در ماتریس قلم-کاربر است. این مشکل از آنجا ناشی می‌شود که کاربران تنها به تعداد کمی از اقلام موجود در پایگاه داده، امتیاز می‌دهند و این امر باعث ناقص‌شدن ماتریس قلم-کاربر می‌شود؛ بنابراین، شباهت محاسبه‌شده بین کاربران/اقلام، غیرقابل اعتماد هستند، زیرا تعدادی از امتیازات همپوشان بین آنها وجود دارد. مشکل دیگری که رویکرد حافظه پایه از آن رنج می‌برد مسئله کارایی است. در این روش برای یافتن همسایگان مشابه نیاز به اندازه‌گیری شباهت بین جفت اقلام یا کاربران است. به‌منظور برطرف‌کردن مشکل یادشده و افزایش کارایی روش حافظه پایه، یک‌سری از پژوهش‌های پیوسته توسط پژوهش‌گران پیشین انجام شده است، آنها برای حل این معضل استفاده از یک رویکرد خوشه‌بندی مدل پایه را پیشنهاد داده‌اند [19-23]. در این روش پیشنهادی، اقلام یا کاربران مشابه را به خوشه‌های جداگانه‌ای، به‌منظور شناسایی کاربران یا اقلام همسایه (که منجر به ارائه پیشنهادهایی به کاربر فعال می‌شود) گروه‌بندی می‌کند. با این حال، استفاده از خوشه‌بندی ممکن است، منجر به برخی کاستی‌هایی، همچون کاهش قابلیت مقیاس‌پذیری، کاهش دقت، هم‌پوشانی[7] و پیشنهادهای بسیار کلی شود؛ که ممکن است توسط محققان نادیده گرفته شود.

دلیل ایجاد کاهش مقیاس‌پذیری در روش بالا که توسط خوشه‌بندی اتفاق می‌افتد، این است که در این روش می‌بایست مقایسه کاربر-کاربر، در هر خوشه مشابه برای شناسایی همسایگان کاربر فعال انجام شود. همچنین دلیل کاهش دقت در روش بالا این است که ممکن است پیشنهادات ارائه شده به کاربر، بر اساس نمایندگانی از خوشه‌های مشابه ارائه شده باشد که ممکن است کاربران یا اقلامی باشند که واقعی نیستند. روش خوشه‌بندی ممکن است منجر به همپوشانی نیز شود؛ زیرا کاربران یا اقلام ممکن است به چند دسته تقسیم شوند. علاوه بر این، روش خوشه‌بندی ممکن است منجر به ارائه پیشنهاداتی بسیار کلی (کمتر شخصی) شود؛ در نتیجه، این مشکل می‌تواند موجب کاهش دقت پیشنهادات ارائه شده به کاربر فعال در مقایسه با روش‌های حافظه پایه گردد.

مدل‌سازی اطلاعات در سطح معنایی، یکی از اهداف اصلی استفاده از هستان‌شناسی است [24]. تعریف اولیه هستان‌شناسی در علوم رایانه توسط گربر[8] در سال ۱۹۹۳ میلادی ارائه شد و بعدها توسط استاب[9] و استادر[10] در سال ۲۰۰۹ تصحیح شد. مفهوم هستان‌شناسی در ابتدا توسط گربر [25] به‌عنوان "توصیف صریح از یک مفهوم" ارائه شده است. بورست[11] [26] هستان‌شناسی را به‌عنوان "توصیف رسمی از یک مفهوم مشترک" تعریف می‌کند. علاوه بر این، تانیار[12] و رحاو[13] [27] هستان‌شناسی را "به‌عنوان دانش مفهوم‌سازی دامنه که به‌وسیله رایانه قابل پردازش است و قادر است واقعیات، ویژگی‌ها و بدیهیات را مدل‌سازی کند" تعریف می‌کند. بر طبق نظر آنتونیو و ون هارلمن [28]

[4] Classification
[5] Vector Space
[6] Ontology-Based

[7] Overlapping
[8] Gruber
[9] Staab
[10] Studer
[11] Borst
[12] Taniar
[13] Rahayu

[DOI: 10.61186/jsdp.20.3.197]

هستان‌شناسی مشخصاً از "یک فهرست لغات و رابطه میان مفاهیم" ساخته شده است.

هستان‌شناسی شامل ویژگی‌های مفهومی، عبارات، محدودیت‌های مربوط به رابطه‌ها، و توصیف روابط منطقی بین اشیاء است. هستان‌شناسی ابزاری است برای ساخت مدل رسمی ساختار یک سامانه، بر اساس روابط حاصل از مشاهدات در آن.

زمانی که هستان‌شناسی تنها شامل روابط IsA است به جای عبارت هستان‌شناسی از "طبقه‌بندی اصطلاح" (سلسله مراتب موضوع) استفاده می‌شود و معمولاً استفاده از کلمه هستان‌شناسی موقعی صحیح است که سامانه مورد بررسی، در برگیرنده انواع روابط متقابل بین مفاهیم (شامل گزاره‌های منطقی که توصیف کننده ارتباط بین مفاهیم هستند) باشد.

الگوریتم‌های پالایش مشارکتی پایه و هستان‌شناسی پایه که در سامانه‌های پیشنهاد شده استفاده می‌شوند، دو شاخص اساسی برای طبقه‌بندی مقاله‌های پژوهشی انجام شده در این حوزه است. مفاهیم هستان‌شناسی به‌عنوان ابزاری جهت تسهیل شناسایی اطلاعات موجود در پروفایل کاربران است، اکثر سامانه‌های پیشنهادگر موفق، سعی می‌کنند به واسطه دانش هستان‌شناسی، دقت پیشنهادات ارائه شده به کاربران را افزایش دهند [29]. بر اساس تحقیقی که توسط میدلتون و همکاران [29] صورت گرفته است، در کنار استفاده از الگوریتم‌های یادگیری ماشین می‌توان سامانه‌های پیشنهادگر را با استفاده از مفاهیم هستان‌شناسی بهبود داد. امروزه، در سامانه‌های پیشنهادگر علاوه بر این‌که از تکنیک‌هایی همچون روش‌های تشخیص آماری الگو، یادگیری ماشین و روش‌های ابتکاری استفاده می‌شود، سعی می‌گردد از مفاهیم هستان‌شناسی نیز جهت بهبود نتایج استفاده شود [30]. در اکثر سامانه‌های پیشنهادگر، معمولاً از تکنیک‌هایی استفاده می‌شود که دارای مزایایی همچون دقت، ارتباط بین اقلام و کالاها، بازخورد و یا قابلیت رصد کردن رفتار کاربر در طول زمان باشند. در این سامانه‌ها، اقلام جدید می‌توانند با استفاده از جستجوی قلم-قلم ، مشابه کاری که در روش پالایش محتوا پایه انجام می‌شود، به کاربران پیشنهاد داده شوند. همچنین ارزیابی یک قلم به‌خصوص برای این‌که مشخص شود آیا برای یک کاربر خاص مناسب هست یا خیر، از طریق بررسی پروفایل جمعی از کاربران، مشابه کاری که در روش پالایش مشارکتی انجام می‌شود، قابل انجام است. همچنین می‌توان از مزایای استفاده از ارتباطات معنایی بین اقلام موجود در پایگاه داده، در روش‌هایی همچون روش پالایش ترکیبی و روش‌های ابتکاری بهره جست [30].

اگر داده‌های آموزشی در دسترس باشد، استفاده از روش پالایش محتوا پایه می‌تواند مؤثر باشد. در مقابل، اگر سامانه دارای تعداد قابل ملاحظه‌ای از کاربران باشد، روش پالایش مشارکتی در مقایسه با روش پالایش محتوا پایه بهتر عمل خواهد کرد. با این وجود، هیچ قاعده و قانونی وجود ندارد که بتوان گفت دقیقاً چه نوع از این روش‌ها را می‌توان مورد استفاده قرار داد [30]. به‌منظور بهبود و توسعه سامانه‌های پیشنهادگر محتوا پایه، می‌توان از تکنیک‌های هستان‌شناسی همچون OntoSeek استفاده کرد [31].

از OntoSeek می‌توان به‌منظور فرموله کردن درخواست‌ها استفاده کرد. همچنین سامانه‌های هستان‌شناسی پایه می‌توانند برای ایجاد (خودکار) پایگاه‌های دانش، از صفحات وب (به‌عنوان مثال Web-KB) استفاده نمایند [32].

در Web-KB نمونه‌هایی از صفحات وب وجود دارند که به‌صورت دستی برچسب‌گذاری شده‌اند و این سامانه قادر است به‌وسیله تکنیک‌های یادگیری ماشین به‌طور خودکار صفحات جدید وب را رده‌بندی کند. این سامانه‌ها اطلاعات پویا و در حال تغییر، مانند علایق کاربران را ذخیره نمی‌کنند. گراف هستان‌شناسی اقلام، می‌تواند ارتباط معنایی اقلام مختلف را نمایان سازد، و این امر می‌تواند موجب افزایش اثربخشی پیشنهادات ارائه شده به کاربران گردد.

روش‌های پروفایل‌سازی هستان‌شناسی پایه، که برای مثال در سامانه‌های Foxtrot و Quickstep استفاده می‌شود [29] با هدف پرکردن شکاف معنایی بین ویژگی‌های سطح پایین استخراج شده از اسناد و مشاهدات مفهومی مورد علاقه کاربران، انجام می‌شود [33]. در واقع، مفاهیم هستان‌شناسی می‌تواند برای بهبود عملکردِ پروفایل کاربر در سامانه‌های پیشنهادگر توسعه یافته، استفاده شوند.

بسط و گسترش مجموعه لغات با استفاده از هستان‌شناسی، یکی از روش‌هایی است که می‌توان از آن برای پرکردن شکاف معنایی بین مفاهیم استفاده‌شده در پروفایل کاربران و حاشیه نویسی‌های مربوط به تصاویر اقلام، از آن استفاده کرد. به‌طورمعمول از هستان‌شناسی‌های دامنه‌ای برای اتصال مفاهیم، بین پروفایل کاربران و اقلام (از طریق ساختار سلسله‌مراتبی) استفاده می‌شود [29]. بر مبنای پژوهشی که توسط گاچ و همکاران [34] انجام شده است، می‌توان از طریق ارزیابی رفتار یک کاربر خاص، به‌وسیله اندازه‌گیری پارامترهایی همچون محتوا و زمان صرف شده در هر صفحه وب، پروفایل کاربر مورد نظر را بهبود بخشید.

[DOI: 10.61186/jsdp.20.3.197]

در سامانه‌های پیشنهادگر، اطلاعات معنایی مربوط به یک قلم به‌خصوص، شامل مواردی همچون: ویژگی‌ها، روابط میان اقلام و همچنین رابطه بین اطلاعات غیر نمادین[14] و اقلام می‌شود. در سال‌های اخیر، تکنیک‌های هستان‌شناسی به‌طور موفقیت آمیزی در سامانه‌های پیشنهادگر برای غلبه بر نقص‌های این سامانه‌ها مورد استفاده قرار گرفته‌اند [5,35]. پورسل و همکاران سعی کردند در روش پیشنهادی خود با استفاده از هستان‌شناسی فازی، دقت سامانه‌های پیشنهادگر را بهبود بخشند [11]. بسیاری از پژوهشگران در حوزه سامانه‌های پیشنهادگر به‌منظور اندازه‌گیری میزان علاقه کاربران به ویژگی‌های محتوا پایه اقلام، از تکنیک‌های هستان‌شناسی استفاده کرده‌اند [30,36,37]. علاوه بر این، همان‌طور که می‌دانیم مشکل شروع سرد که در نتیجه ورود قلم (یا اقلام) جدید و یا کاربر جدید به سامانه اتفاق می‌افتد یکی از چالش‌های اساسی در سامانه پیشنهادگر است، به همین خاطر بسیاری از پژوهش‌گران پیشنهاد می‌کنند برای حل این مشکل، تکنیک‌های هستان‌شناسی را با روش‌هایی همچون پالایش مشارکتی و محتوا پایه ترکیب کنیم [30,38-41].

همچنین، برخی از پژوهش‌گران پیشنهاد می‌کنند، به‌منظور بهبود سامانه‌های پیشنهادگر مفاهیم پایه، می‌توان از ترکیب روش پالایش مشارکتی (قلم پایه) با تکنیک‌های شباهت معنایی (همچون هستان‌شناسی) استفاده کرد [42]. راه‌کارهای مطرح‌شده در بالا می‌تواند برای تولید پیشنهادها در سامانه‌ها و حوزه‌هایی که برمبنای روابط معنایی و دانش، کار می‌کنند، مورد استفاده قرار گیرد (به‌ویژه در برنامه‌های کاربردی وب۲) [44, 43].

هستان‌شناسی، نوعی مفهوم‌سازی دامنه‌ای است که قابل خواندن برای ماشین نیز است. هستان‌شناسی به‌طورمعمول به‌صورت ساختاری است که روابط میان مفاهیم، عناصر و ویژگی‌ها را نشان می‌دهد [12]. پیدا کردن شباهت معنایی بین مفاهیم و در کل هستان‌شناسی به‌عنوان یک ساختار پیچیده (به‌عنوان مثال Flickr) اهمیت دارد [45].

۳- روش پیشنهادی

در این بخش از مقاله سعی خواهد شد به نکات روش پیشنهادی مشتمل بر مفاهیم مورد نیاز و مراحل اجرای آن پرداخته شود.

سامانه پیشنهادگر ارائه‌شده در این مقاله، مبتنی بر روش ترکیبی است. در روش ترکیبی پیشنهاد شده به‌منظور به‌دست‌آوردن نتیجه مطلوب، از دو روش پالایش محتوا پایه و پالایش مشارکتی که خود ترکیبی از روش‌های حافظه پایه و مدل پایه می‌باشد، استفاده شده است.

در بخش پالایش مشارکتی چندین تکنیک از قبیل خوشه‌بندی، استفاده از اطلاعات پروفایل کاربر با کمک تکنیک‌های هستان‌شناسی، هستان‌شناسی اقلام، به‌کار بردن تکنیک شباهت معنایی در هستان‌شناسی برای غلبه بر مشکلاتی همچون پراکندگی داده‌ها، مسئله شروع سرد، کندی و کمبود کیفیت پیشنهادات ارائه شده، استفاده می‌شود. همچنین در بخش پالایش محتوا پایه‌ی روش پیشنهادی، از تکنیک هستان‌شناسی قلم پایه و شباهت معنایی استفاده می‌شود. لازم به ذکر است، در هنگام استفاده از تکنیک شباهت معنایی، به‌منظور افزایش دقت در اندازه‌گیری میزان شباهت‌های یال‌های بین دو قلم شاخص (IsA)، از یک روش ابتکاری به‌منظور ارائه پیشنهادهای دقیق‌تر به کاربر فعال، استفاده می‌شود.

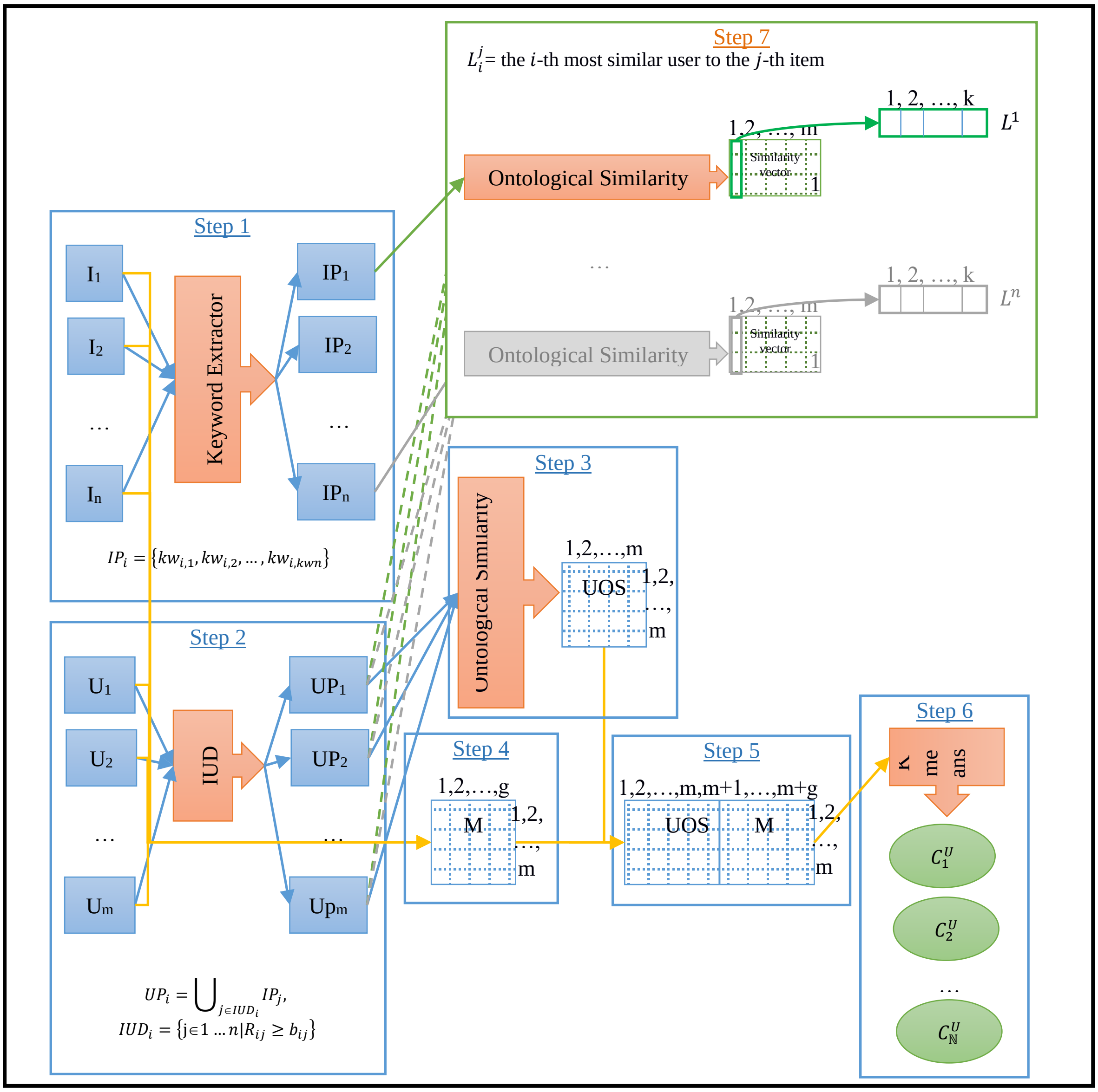
در شکل‌های ۱ و ۲، ساختار کلی[15] سامانه پیشنهادگر ترکیبی نشان داده شده است. همان‌طور که در این شکل‌ها دیده می‌شود، روش ترکیبی پیشنهادشده شامل سه بخش اصلی است: پالایش محتوا پایه و پالایش مشارکتی که ترکیبی از روش‌های حافظه پایه و مدل پایه است. در روش پیشنهادی، بخش پالایش مشارکتی از یک‌سو، از مخزن هستان‌شناسی فیلم‌ها[16] و اطلاعات ضمنی کاربران[17] برای ساخت هستان‌شناسی پروفایل کاربران استفاده می‌کند و از سوی دیگر، از اطلاعات رتبه‌بندی صریح[18] کاربران به‌عنوان دانش مفید در مرحله خوشه‌بندی استفاده می‌شود و هستان‌شناسی را تکمیل می‌کند.

در بخش پالایش محتوا پایه از دانش موجود در مخزن هستان‌شناسی فیلم‌ها جهت مشخص‌کردن میزان "درجه IsA"های استفاده‌شده در هستان‌شناسی استفاده می‌شود که در نتیجه منجر به وزن‌دارشدن مفاهیم موجود در درخت سلسله‌مراتبی هستان‌شناسی می‌شود و به این ترتیب می‌توان بهترین اقلام را برای پیشنهاد به کاربر هدف، مشخص کرد. همان‌طور که ملاحظه می‌شود، شکل‌های (۱) و (۲) نشان می‌دهد که اندازه‌گیری "درجه IsA"ها، قبل از اندازه‌گیری میزان شباهت معنایی، یکی از مراحل اصلی در روش ترکیبی پیشنهادی است.

[14] Meta-Information

[15] Framework
[16] Movies Ontology Repository
[17] Implicit Users Information
[18] Explicit Rating

[DOI: 10.61186/jsdp.20.3.197]



(شکل-۱): مرحله آموزش سامانه پیشنهادی

Figure 1: Training phase of the proposed system

خوشه‌بندی، هستان‌شناسی پروفایل کاربر، و الگوریتم پیشنهادی حافظه پایه (الگوریتم *k-NN* بهبود یافته) در بخش پالایش مشارکتی استفاده می‌شوند. همچنین شبکه‌کد روش پیشنهادی نیز در بخش ضمائم ارائه شده است.

در ادامه، فرایندی که در هر یک از مراحل اجرای سامانه پیشنهادی رخ می‌دهد و این‌که چه مؤلفه‌هایی در هر فرایند درگیر هستند، توضیح داده می‌شود.

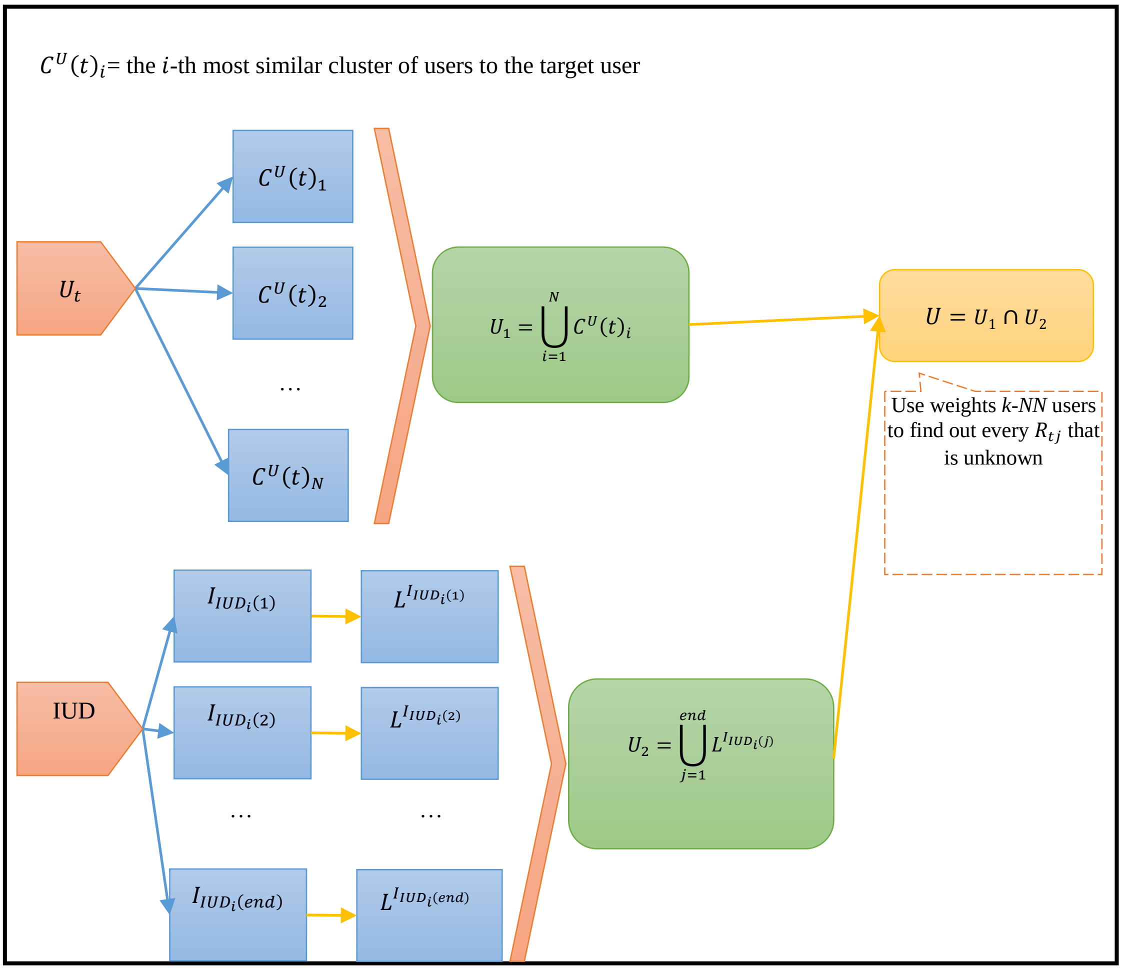
سه بخش در روش پیشنهادی وجود دارد. دو بخش نخست، دو سامانه پیش‌بینی مقادیر گم‌شده به کمک دو رویکرد و بخش سوم ترکیب‌گر[19] نتایج بخش قبل است. بخش نخست و دوم در دو کد، الگوریتم (۱) و الگوریتم (۲) ارائه شده است؛ و بخش سوم نیز در کد الگوریتم (۳) ارائه شده است. الگوریتم (۱) به یک روش پرکردن مقادیر گم شده در حالتی که مجموعه‌داده بسیار تُنُک باشد پرداخته است. این روش که بر اساس خوشه‌بندی مبتنی بر فاصله ی "گری" است را ابتدا توضیح می‌دهیم.

نظریه سامانه‌های گری[20] [46] برای سامانه‌های نامشخص و ناقص مناسب است و توانایی استخراج اطلاعات پنهان از اطلاعات موجود را داراست [47]. برنامه‌های کاربردی (واقعی) زیادی [48-50] با موفقیت از آن استفاده کرده‌اند. فاصله‌ی بین هر جفت نمونه[21] قابل اندازه‌گیری است.

[19] Aggregator

[20] Grey System Theory
[21] Pair Of Samples

[DOI: 10.61186/jsdp.20.3.197]



(شکل – ۲): مرحله آزمایش سامانه پیشنهادی

Figure 2: Test phase of the proposed method

فاصله بین یک جفت نمونه (در اینجا، یک نمونه می‌تواند یک کاربر باشد) را می‌توان با در نظر گرفتن کل مجموعه‌داده (در اینجا مجموعه‌داده ، ماتریس رتبه‌بندی R است) با توجه به رابطه زیر محاسبه کرد.

$$F_{ij}(R) = \frac{\sum_{q=1}^{\alpha} f_{ijq}(R)}{\alpha} \qquad (۱)$$

که:

$$f_{ijq}(R) = 1 - \frac{\min_{I\in\{1,\dots,\alpha\}} \min_{J\in\{1,\dots,\beta\}} |R_{iJ} - R_{IJ}| + \theta \max_{I\in\{1,\dots,\alpha\}} \max_{J\in\{1,\dots,\beta\}} |R_{iJ} - R_{IJ}|}{|R_{iq} - R_{jq}| + \theta \max_{I\in\{1,\dots,\alpha\}} \max_{J\in\{1,\dots,\beta\}} |R_{iJ} - R_{IJ}|} \qquad (۲)$$

در رابطه بالا، θ به‌عنوان یک عدد حقیقی در بازه 0 تا 1 بوده که معمولاً مقدار آن برابر 0.5 است [51]. اصولاً $F_{ij}(R)$ به صورت یک عدد حقیقی در بازه 0 تا 1 قرار دارد و قابل اثبات است. اگرچه تابع F یک معیار دقیق فاصله نیست، اما می‌توان آن را یک تابع فاصله تقریبی در نظر گرفت [52]. سپس این روند را به یک مجموعه‌داده کامل تعمیم می‌دهیم، نمونه تعمیم داده شده فاصله گری برای یک مجموعه‌داده ناقص در رابطه زیر ارائه شده است.

$$\tilde{F}_{ij}(R) = \frac{\sum_{q=1}^{\alpha} \Phi_{ijq}(R) \times \tilde{f}_{ijq}(R)}{\sum_{q=1}^{\alpha} \Phi_{ijq}(R)} \qquad (۳)$$

که:

$$\tilde{f}_{ijq}(R) = \Phi_{ijq}(R) \times \left[1 - \frac{\min_{I\in\{1,\dots,\alpha\}} \min_{J\in\{1,\dots,\beta\}} |R_{iJ} - R_{IJ}| + \theta \max_{I\in\{1,\dots,\alpha\}} \max_{J\in\{1,\dots,\beta\}} |R_{iJ} - R_{IJ}|}{|R_{iq} - R_{jq}| + \theta \max_{I\in\{1,\dots,\alpha\}} \max_{J\in\{1,\dots,\beta\}} |R_{iJ} - R_{IJ}|}\right] \qquad (۴)$$

$$\Phi_{ijq}(R) = \begin{cases} 1 & R_{iq} \neq NaN \;\&\; R_{jq} \neq NaN \\ 0 & o.w. \end{cases} \qquad (۵)$$

همانطور که مشاهده می‌کنید این معیارِ عدم تشابه، در مجموعه‌داده های ناقص بهتر از معیار مینکوفسکی[۲۲] عمل می‌کند. همچنین، در مجموعه‌داده های ناقص نسبت به معیار فاصله اقلیدسی[۲۳] برتری دارد [47]. با این حال، طبق تحقیق [53] که در آن فاصله اقلیدسی وزن‌دار[۲۴] معرفی و استفاده شده است، ما اندازه‌گیری عدم تشابه خود را طبق رابطه زیر تعمیم می‌دهیم.

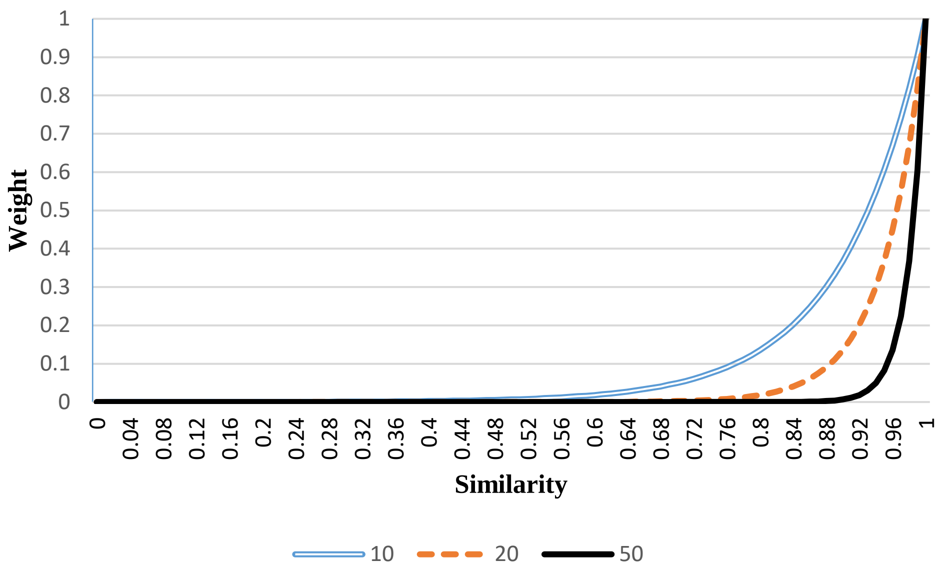
$$\mathcal{D}_{ij}(R) = \sqrt{\frac{\alpha}{\sum_{q=1}^{\alpha} \Phi_{ijq}(R)}} \times \tilde{F}_{ij}(R) \qquad (۶)$$

که:

$$\mathcal{S}_{ij}(R) = \sqrt{\frac{\sum_{q=1}^{\alpha} \Phi_{ijq}(R)}{\alpha}} \times \left(1 - \tilde{F}_{ij}(R)\right) \qquad (۷)$$

[22] Minkowski
[23] Euclidean Distance
[24] Weighted Euclidean Distance

[DOI: 10.61186/jsdp.20.3.197]



(شکل- ۳): تأثیر مقادیر مختلف پارامتر b بر تابع وزن؛ در اینجا، bروی ۱۰، ۲۰ و ۵۰ تنظیم شده است

Figure 3: Effect of different values of parameter b on weight function; here, b is set to 10, 20, and 50.

برای تخمین مقدار رتبه‌بندی گمشده[1]، از سازوکار وزنی استفاده می کنیم؛ که در آن برای هر کاربرِ کاندیدی که جهت شرکت در محاسبه مقدار رتبه‌بندی، انتخاب می‌شود باید وزنی در نظر گرفته شود. وزن هر کاربر منتخب بر اساس شباهت آن با کاربر هدف تعریف شده و با استفاده از رابطه زیر محاسبه می‌شود.

$$w_q = \frac{e^{b s_{tq}(R)}}{e^{b}} \qquad (۸)$$

در رابطه بالا، t به‌عنوان شاخص کاربر هدف، q به‌عنوان شاخص کاربر منتخب و b پارامتر تنظیم [54] برای تبدیل شباهت به وزن است. تأثیرات پارامترهای یادشده را می‌توان در شکل (۳) مشاهده کرد. بعد از چندین دوره آزمایشی، مقدار ۳۰ را برای این پارامتر انتخاب می‌کنیم.

الگوریتم (۲) که توصیف کننده بخش دوم هست یک شبه‌کد سامانه پیشنهادگر ترکیبی هستان‌شناسی پایه است. شبه‌کد سامانه پیشنهادگر ترکیبی هستان‌شناسی پایه که شامل استخراج‌کننده پروفایل کاربران و اقلام[2] است، در الگوریتم (۲) (رجوع شود به ضمایم) ارائه شده‌است. همان‌طور که مشاهده می‌کنید، این الگوریتم، ماتریس رتبه‌بندی (که ممکن است ناقص باشد) و متن توضیحات مربوط به هر قلم و همچنین تعداد کلمات کلیدی را به‌عنوان پارامترهای ورودی می‌گیرد و سپس پروفایل‌های کاربران و اقلام را برمی‌گرداند. این کار بر اساس مفهوم پالایش مشارکتی کاربر پایه در صورت وفور داده‌های لازم و عدم سردبودن کاربر هدف، یا بر اساس پالایش مشارکتی قلم پایه در صورت وفور داده‌های لازم و عدم سردبودن قلم هدف، عمل می‌کند. در غیر این صورت‌ها بر اساس مفهوم هستان‌شناسی، خوشه‌بندی-کاربر و پالایش مبتنی بر محتوای TF-IDF پایه، عمل می‌کند.

الگوریتم (۳) (رجوع شود به ضمایم) مربوط به شبه‌کد روش پیشنهادی است. همان‌طور که مشاهده می‌شود، این الگوریتم، ماتریس رتبه‌بندی (که یک ماتریس ناقص است) و تعداد خوشه‌ها را به‌عنوان پارامترهای ورودی می‌گیرد و سپس یک نسخه کامل از ماتریس رتبه‌بندی و خوشه‌ای از کاربران را برمی‌گرداند. جمع‌آوری داده‌ها، پالایش مشارکتی و پالایش محتوا پایه، و روش ترکیبی، سه بخش از این سامانه می‌باشند.

در بخش جمع‌آوری داده‌ها، سه فرایند شامل آماده-سازی داده‌ها، ساخت هستان‌شناسی اقلام، و ساخت هستان‌شناسی پروفایل کاربران انجام می‌شود. در بخش دوم، پالایش مشارکتی و پالایش محتوا پایه می‌توانند به‌صورت موازی انجام شوند.

فرض کنید یک ماتریس R از امتیازدهی‌های گوناگون داریم. یکی از راه‌ها، کاهش ابعاد این ماتریس به روش فاکتورسازی ماتریس [55]، روش SVD [56] و یا سایر روش‌های کاهش بعد مرتبط [57-59] است. متعارف‌ترین روش پیشنهادگر در سامانه‌های مشارکتی، استفاده از k نزدیک‌ترین اقلام و یا کاربر همسایه برای پیش‌بینی مقادیر ناموجود مورد نظر است. برای مثال در روش k نزدیک‌ترین اقلام همسایه (قلم پایه)، مقادیر گم‌شده را با استفاده از رابطه زیر پیش‌بینی می‌کنیم.

[1] Missing Rating Value
[2] Users-Items Profile Extractor

[DOI: 10.61186/jsdp.20.3.197]

$${}_I\hat{R}_{ij} = b_{ij} + \frac{\sum_{j'\in NNI_j^{ik}} \pi\left({}_I s_{jj'}, type\right)\left(R_{ij'} - b_{ij'}\right)}{\sum_{j'\in NNI_j^{ik}} \pi\left({}_I s_{jj'}, type\right)} \qquad (۹)$$

که R ماتریس امتیازدهی، ${}_I\hat{R}$ تخمین ماتریس امتیازدهی با کمک روش قلم پایه، R_{ij} امتیازی که کاربر i-ام به قلم j-ام می‌دهد، ${}_I\hat{R}_{ij}$ تخمین امتیازی که کاربر i-ام به قلم j-ام می‌دهد با کمک روش قلم پایه، NNI_j^{ik} یک مجموعه k عضوی از اندیس اقلامی که بیشترین شباهت را با به قلم j-ام دارند و توسط کاربر i-ام امتیازدهی شده‌اند، b_{ij} پایه امتیازدهی کاربر i-ام به علاوه پایه امتیازدهی قلم j-ام و پایه امتیازدهی سامانه، $type$ نشان‌گر نوع وزن‌دار یا غیر وزن‌دار روش و در نهایت ${}_I s_{ij}$ میزان شباهت بین اقلام i-ام و j-ام است. در اینجا و در تمام روابط $\pi\left(R_{ij}, Crisp\right)$ از رابطه زیر محاسبه می‌شود:

$$\pi\left(R_{ij}, Crisp\right) = \begin{cases} 0 & R_{ij} = NaN \\ \frac{R_{ij}}{R_{ij}} & Crisp = 1 \;\&\; R_{ij} \neq NaN \\ R_{ij} & Crisp = 0 \;\&\; R_{ij} \neq NaN \end{cases} \qquad (۱۰)$$

همچنین b_{ij} از رابطه زیر به دست می‌آید:

$$b_{ij} = b_{i-} + b_{-j} + b \qquad (۱۱)$$

که b پایه امتیازدهی سامانه، b_{i-} پایه امتیازدهی کاربر i-ام، b_{-j} پایه امتیازدهی قلم j-ام است. b پایه امتیازدهی سامانه برابر با میانگین ماتریس امتیازدهی یعنی $\bar{R}$ است که از رابطه زیر به دست می‌آید:

$$b = \bar{R} = \frac{\sum_i \sum_j \pi\left(R_{ij}, 0\right)}{\sum_i \sum_j \pi\left(R_{ij}, 1\right)} \qquad (۱۲)$$

b_{-j} پایه امتیازدهی قلم j-ام از رابطه زیر به دست می‌آید:

$$b_{-j} = \frac{\sum_i \pi\left(R_{ij} - b, 0\right)}{\lambda_1 + \sum_i \pi\left(R_{ij}, 1\right)} \qquad (۱۳)$$

که λ_1 یک پارامتر منظم‌ساز قابل تنظیم با آزمایش است. گزاره b_{i-} نشانگر پایه امتیازدهی کاربر i-ام است که از رابطه زیر به دست می‌آید.

$$b_{i-} = \frac{\sum_j \pi\left(R_{ij} - b - b_{-j}, 0\right)}{\lambda_2 + \sum_j \pi\left(R_{ij}, 1\right)} \qquad (۱۴)$$

که λ_2 یک پارامتر منظم‌ساز قابل تنظیم با آزمایش است.

بدون ازدست‌دادن عمومیت، با ترانهاده کردن ماتریس R متعارف‌ترین روش پیشنهادگر در سامانه‌های مشارکتی استفاده از k نزدیکترین کاربر همسایه برای پیش‌بینی مقادیر ناموجود در روابط بالا تعریف می‌شود. مثلاً در روش k نزدیکترین کاربر همسایه (کاربر پایه)، مقادیر گم‌شده را با استفاده از رابطه زیر پیش‌بینی می‌کنیم:

$${}_U\hat{R}_{ij} = b_{ij} + \frac{\sum_{i'\in NNU_i^{jk}} \pi\left({}_U s_{ii'}, type\right)\left(R_{i'j} - b_{i'j}\right)}{\sum_{i'\in NNU_i^{jk}} \pi\left({}_U s_{ii'}, type\right)} \qquad (۱۵)$$

که در اینجا NNU_i^{jk} یک مجموعه k عضوی از اندیس کاربرانی است که بیشترین شباهت را به کاربر i-ام دارند و به قلم j-ام امتیازدهی کرده‌اند و ${}_U s_{ij}$ میزان مشابهت کاربران i-ام و j-ام است.

در نمایش متنی هر قلم، اقلام به شکل یک فضای ماتریسی $TFIDF$ نشان داده می‌شود درایه $TFIDF_{i\omega}$ از این ماتریس نشان‌دهنده اهمیت ویژگی یا کلمه ω در متن یا قلم i است که از رابطه زیر به دست می‌آید.

$$TFIDF_{i\omega} = tf_{i\omega} \times idf_{i\omega} \qquad (۱۶)$$

که $tf_{i\omega}$ تعداد رخداد کلمه ω در متن یا قلم i است و $idf_{i\omega}$ برابر با $\log\frac{a}{a_\omega}$ است که a تعداد اقلام و a_ω تعداد اقلامی که حاوی کلمه ω است. می‌توان با استفاده از تکنیک‌هایی شبیه به مقاله [60]، این ماتریس را ابتدا به عناصر SVD، یعنی به ماتریس‌های U، V و Σ، تفکیک کنیم. ماتریس‌های U و V، ماتریس‌های متعامد و ماتریس Σ، ماتریسی قطری است که حاوی مقادیر تکین (Singular Value) در این تفکیک است. با حفظ δ-بزرگترین مقادیر در ماتریس قطری Σ و جایگزینی بقیه عناصر با صفر، و سپس انجام عملیات معکوس، ماتریس ویژگی‌های $TFIDF_{i\omega}$ جدیدی به دست می‌آید که فقط δ ویژگی بامعنی دارد. پس چنان‌که تعداد کل کلمات زیاد باشد و این ماتریس حجیم باشد، با این کار می‌توانیم تعداد کلمات این ماتریس را به شکل کنترل‌شده‌ای کاهش دهیم. حال ماتریس مشابهت قلم-قلم ${}_I s_{ij}^{TFIDF}$ را به شکل زیر تعریف می‌کنیم.

$${}_I s_{ij}^{TFIDF} = \frac{\sum_{\omega\in W}\left[TFIDF_{i\omega} \times TFIDF_{j\omega}\right]}{\sqrt{\sum_{\omega\in W}\left[TFIDF_{i\omega} \times TFIDF_{i\omega}\right]}\sqrt{\sum_{\omega\in W}\left[TFIDF_{j\omega} \times TFIDF_{j\omega}\right]}} \qquad (۱۷)$$

که W مجموعه لغات مورد استفاده در متون یا اقلام یا در صورت استفاده از روش مقاله [60]، تعداد کنترل شده آنها (یعنی δ) است.

علاوه بر روش بالا که عموماً اجازه می‌دهیم تا حجم لغات (یعنی $|W|$) نامحدود باشند، روش‌های دیگری وجود دارند که این کار را بر روی یک مجموعه مشخص از W انجام می‌دهند. این مجموعه W می‌تواند از پیش تعیین شده باشد [61] یا در حین سازوکار سامانه تولید شوند [62]؛ سپس از رابطه بالا استفاده می‌کنیم تا مشابهت اقلام را اندازه‌گیری کنیم. از آنجایی که در این حالت فراوانی کلمات استفاده نمی‌شود، $tf_{i\omega}$ بی‌معنی است، و $TFIDF$ را

با $LimitedIDF$ در رابطه بالا جایگزین کرده و رابطه زیر را به‌دست می‌آوریم:

$$ {}_{I}S_{ij}^{LimitedIDF} = \frac{\sum_{\omega\ shared\ in\ i\text{th}\ and\ j\text{th}\ Items}\left[\frac{a}{a_\omega}\times\frac{a}{a_\omega}\right]}{\sqrt{\sum_{\omega in\ i\text{th}\ Item}\left[\frac{a}{a_\omega}\right]\times\sum_{\omega\ in\ j\text{th}\ Item}\left[\frac{a}{a_\omega}\right]}} \qquad (۱۸)$$

در این حالت که کلمات محدود هستند، می‌توانیم ویژگی باینری $Bin_{i\omega}$ را که نشانگر وجود کلمه ω در متن یا قلم i است، به‌صورت زیر تعریف کنیم:

$$Bin_{i\omega} = \begin{cases} 1 & \omega\ is\ in\ i\text{th}\ Item \\ 0 & \text{Otherwise} \end{cases} \qquad (۱۹)$$

شباهت قلم-قلم ${}_{I}S_{ij}^{JACCARD}$ به‌صورت زیر تعریف می‌شود:

$$ {}_{I}S_{ij}^{JACCARD} = \frac{\sum_{\omega\in W}\left[Bin_{i\omega}\times Bin_{j\omega}\right]}{2\times|W| - \sum_{\omega\in W}\left[Bin_{i\omega}\times Bin_{j\omega}\right]} \qquad (۲۰)$$

معیار شباهت قلم-قلم ${}_{I}S_{ij}^{DICE}$ به‌صورت زیر تعریف می‌شود:

$$ {}_{I}S_{ij}^{DICE} = \frac{\sum_{\omega\in W}\left[Bin_{i\omega}\times Bin_{j\omega}\right]}{2\times\left(\sum_{\omega\in W}\left[Bin_{j\omega}\right]+\sum_{\omega\in W}\left[Bin_{i\omega}\right]\right)} \qquad (۲۱)$$

در یک سامانه مبتنی بر مشابهت تقویت شده با هستان‌شناسی، ابتدا باید یک هستان‌شناسی را به‌صورت رسمی تعریف کنیم [63]. یک هستان‌شناسی $O \equiv (E,\theta,\mathfrak{I})$ شامل یک مجموعه از موجودیت‌ها می‌باشد که با E نمایش داده می‌شوند؛ به‌گونه‌ای که یک گره ریشه به شکل θ می‌باشد (یعنی $\theta \in E$). یک هستان‌شناسی همچنین یک تابع $\mathfrak{I}$ که یک عضو از E را به عضوی دیگر از E نگاشت می‌دهد؛ به‌طوری که موجودیت $\mathfrak{I}(i)$ والد موجودیت i محسوب می‌شود و به معنی آن است که موجودیت i نوعی از موجودیت $\mathfrak{I}(i)$ است و به‌عبارتی $i\ \ IsA\ \mathfrak{I}(i)$. می‌توان یک هستان‌شناسی را به شکل یک درخت با ریشه θ و یال‌های جهت‌داری که هر یالِ جهت‌دار نشانگر یک رابطه IsA است از گره پایین‌تر به گره بالاتر، نشان داد. حال IaD را بین دو موجودیت t_1 و t_2 به‌صورت زیر تعریف می‌کنیم.

$$IaD^{PathDis}(t_1,t_2) = \frac{1}{PathDis(t_1,t_2)+1} \qquad (۲۲)$$

که $PathDis(t_1,t_2)$ کمینه فاصله بین دو موجودیت t_1 و t_2 در درخت هستان‌شناسی بر حسب گام است. حال مشابهت مبتنی بر هستان‌شناسی بین دو قلم یا متن را به شکل زیر تعریف می‌کنیم که p در اینجا یک پارامتر است [63].

$$ {}_{I}^{O}S_{ij}^{X}(p) = \frac{\left\|E_i\cap E_j\right\|_{p,X}}{\sqrt{\left\|E_i\right\|_{p,X}\times\left\|E_j\right\|_{p,X}}} \qquad (۲۳)$$

که E_i مجموعه‌ای از موجودیت‌های موجود در متن یا قلم i-ام است و $\|E_i\|_{p,X}$ میزان مشابهت (از جنس مشابهت IaD^X) درون مجموعه‌اِی مجموعه موجودیت‌های متن یا قلم i-ام (یعنی E_i) می‌باشد و p یک پارامتر است. مقدار $\left\|E_i\cap E_j\right\|_{p,X}$ را برای تعمیم بهتر به صورت زیر تعریف می‌کنیم.

$$\left\|E_i\cap E_j\right\|_{p,X} = \left\|E_i\right\|_{p,X} + \left\|E_j\right\|_{p,X} - \left\|E_i\cup E_j\right\|_{p,X} \qquad (۲۴)$$

نحوه محاسبه مقدار $\|E_i\|_{p,X}$ را به صورت زیر تعریف می‌کنیم.

$$\|E_i\|_{p,X} = \sum_{t_1\in E_i}\left(\frac{1}{\sum_{t_2\in E_i}\left(IaD^X(t_1,t_2)\right)^p}\right) \qquad (۲۵)$$

حال چندین تابع IaD را برای مشابهت‌سنجی دو موجودیت t_1 و t_2 در ادامه تعریف می‌کنیم [63]. ابتدا $IaD^{LCH}(t_1,t_2)$ از رابطه زیر محاسبه می‌شود.

$$IaD^{LCH}(t_1,t_2) = \frac{\log\left(\frac{2d}{PathDis(t_1,t_2)}\right)}{\varepsilon_1} \qquad (۲۶)$$

که d عمق درخت هستان‌شناسی و ε_1 یک پارامتر است. کاربرد این پارامتر در این است که مقدار برد تابع $IaD^{LCH}(t_1,t_2)$ را زیر یک نگاشت کند. پارامتر ε_1 به‌صورت پیش‌فرض ۳ است. با این مقدار تا حدود بسیار خوبی، یک، نشان‌گر مشابهت دو موجودیت t_1 و t_2 است و صفر نشانگر تفاوت آنها می‌باشد. $IaD^{RES}(t_1,t_2)$ از رابطه زیر محاسبه می‌شود.

$$IaD^{RES}(t_1,t_2) = \frac{Info_{NSP_{t_1,t_2}}}{\varepsilon_2} \qquad (۲۷)$$

که NSP_{t_1,t_2} نخستین والد مشترک بین دو موجودیت t_1 و t_2 در درخت هستان‌شناسی، ε_2 یک پارامتر و $Info_t$ که میزان ارزش اطلاعاتی موجودیت t است، از رابطه زیر به‌دست می‌آید. پارامتر ε_2 به‌صورت پیش‌فرض ۱۰ می‌باشد:

$$Info_t = \log\frac{1}{p_t} \qquad (۲۸)$$

که p_t برابر احتمال رخداد موجودیت t در کل اقلام است و از رابطه زیر به‌دست می‌آید:

$$p_t = \frac{\#\ of\ times\ t\ or\ any\ of\ parents\ of\ t\ occurs\ in\ all\ items}{\#\ of\ all\ entities} \qquad (۲۹)$$

$IaD^{JCN}(t_1,t_2)$ از رابطه زیر محاسبه می‌شود:

$$IaD^{JCN}(t_1,t_2) = \frac{1}{\varepsilon_3\times\left(Info_{t_1}+Info_{t_2}-2\times Info_{NSP_{t_1,t_2}}\right)} \qquad (۳۰)$$

که ε_3 یک پارامتر است که به‌صورت پیش‌فرض ۲ است. $IaD^{LIN}(t_1,t_2)$ از رابطه زیر محاسبه می‌شود:

$$IaD^{LIN}(t_1,t_2) = \frac{2\times Info_{NSP_{t_1,t_2}}}{Info_{t_1}+Info_{t_2}} \qquad (۳۱)$$

[DOI: 10.61186/jsdp.20.3.197]

در نهایت $IaD^{WUP}(t_1,t_2)$ از رابطه زیر محاسبه می‌شود:

$$IaD^{WUP}(t_1,t_2)=\frac{2\times d_{NSP_{t_1,t_2}}}{d_{t_1}+d_{t_2}} \qquad (۳۲)$$

که در اینجا d_t نشان‌گر عمق موجودیت t در درخت هستان‌شناسی است. توجه کنید که عمق ریشه درخت هستان‌شناسی، یعنی d_θ، صفر است.

حال رویکرد پیشنهادی برای اندازه‌گیری میزان شباهت معنایی به کمک معیار پیشنهادی به نام IaD^M را شرح می‌دهیم:

$$IaD^M(t_1,t_2,\beta_1,\beta_2)= \begin{cases}\dfrac{\ddot{L}_{t_1t_2}{}^{-1}\times\sum_{d=1}^{\beta_2}Bool\left(t_1\in SChCSet\left(t_2|RD_k^{t_2}\right)\right)}{\sum_{\ddot{L}=1}^{\beta_1}\left[\ddot{L}^{-1}\times\sum_{t'\in ChCSet(t_2,\ddot{L})}\sum_{d=1}^{\beta_2}Bool\left(t'\in SChCSet\left(t_2|RD_k^{t_2}\right)\right)\right]}+\ddot{L}_{t_1t_2}{}^{-1} & t_1\in ChCSet\left(t_2,\ddot{L}_{t_1t_2}\right),\ddot{L}_{t_1t_2}\leq\beta \\ 0 & Otherwise\end{cases} \qquad (۳۳)$$

که RD_k^t نشان‌گر k-امین مرتبط‌ترین قلم یا متن به موجودیت t، β_1 و β_2 دو پارامتر، $ChCSet(\boldsymbol{t},i)$ مجموعه همه موجودیت‌های فرزند موجودیت t با i عمق پایین‌تر، $\ddot{L}_{t_1t_2}$ تفاوت عمق بین دو موجودیت t_1 و t_2 در درخت هستان‌شناسی، $SChCSet(\boldsymbol{t}|RD_k^{\boldsymbol{t}})$ مجموعه همه موجودیت‌های فرزند موجودیت t در RD_k^t هستند. تابع $ChCSet(\boldsymbol{t},i)$ به‌صورت زیر تعریف می‌شود:

$$ChCSet(\boldsymbol{t},i)=\bigcup_{\boldsymbol{t'}\in ChCSet(\boldsymbol{t},i-1)}ChCSet(\boldsymbol{t'},1) \qquad (۳۴)$$

تابع $Bool(\boldsymbol{Condition})$ نیز به‌صورت زیر تعریف می‌شود.

$$Bool(\boldsymbol{Condition})=\begin{cases}1 & if\ \boldsymbol{Condition}\ is\ true\\ 0 & Otherwise\end{cases} \qquad (۳۵)$$

چنان‌که پیش‌تر گفته شد، طبقه‌بندی مفهومی موجودیت‌ها بر اساس میزان شباهت معنایی میان آنها، یکی از مراحل مهم در فرایند ساخت مدل هستان‌شناسی می‌باشد. این‌کار با توجه به جایگاه هر موجودیت خاص در درخت سلسله‌مراتبی قابل انجام است. به‌طوراساسی، در فرایند طبقه‌بندی مفهومی می‌توان گفت مفاهیمی که در درخت سلسله‌مراتبی به یکدیگر نزدیک‌ترند، نسبت به‌هم شبیه‌تر هستند. با این حال، برای محاسبه دقیق میزان شباهت بین مفاهیم، لازم است وزن یال‌های بین دو موجودیت مختلف، محاسبه شود. زیرا شباهت معنایی در طبقه‌بندی مفهومی، بر اساس وزن یال‌ها اندازه‌گیری می‌شود.

برای اندازه‌گیری شباهت معنایی بین دو عبارت در درخت سلسله‌مراتبی وزن‌دار (یال‌های بین موجودیت‌ها وزن‌دار شده‌اند)، نخستین گام، محاسبه فاصله بین دو عبارت است که این کار با کمک وزن‌های به‌دست‌آمده در مراحل قبلی قابل انجام است. فاصله بین دو عبارت با استفاده از رابطه زیر محاسبه می‌شود:

$$Dis(t_1,t_2)=\left[IaD^M(t_1,t_2,\beta_1,\beta_2)\right]^{-1} \qquad (۳۶)$$

که در آن $Dis(t_1,t_2)$ نشان دهنده فاصله بین دو موجودیت یا مفهوم t_1 و t_2 است. میزان تشابه معنایی بین دو مفهوم t_1 و t_2 را می‌توان با استفاده از رابطه زیر استخراج کرد:

$${}_{I}^{O}s_{ij}^{M}(t_1,t_2)=\frac{2\times d_{NSP_{t_1,t_2}}}{d_{t_1}+d_{t_2}-2\times d_{NSP_{t_1,t_2}}}\times V_{t_1,t_2} \qquad (۳۷)$$

که NSP_{t_1,t_2} نخستین والد مشترک بین دو موجودیت t_1 و t_2 در درخت هستان‌شناسی، d_a عمق موجودیت a است. در رابطه بالا V_{t_1,t_2} بر اساس رابطه زیر تعریف می شود:

$$V_{t_1,t_2}=\begin{cases}\dfrac{1}{d_{t_1}-d_{t_2}} & t_1\ is\ an\ ancestor\ of\ t_2\\ \dfrac{1}{d_{t_2}-d_{t_1}} & t_2\ is\ an\ ancestor\ of\ t_1\\ \dfrac{1}{d_{t_1}+d_{t_2}-2\times d_{NSP_{t_1,t_2}}} & Otherwise\end{cases} \qquad (۳۸)$$

این رابطه نشان می‌دهد که واریانس‌های معنایی در بین سطوح بالای سلسله‌مراتب، بیشتر از واریانس‌های معنایی در بین سطوح پایین‌تر است. در این بین، فاصله‌ی بین مفاهیم خواهر و برادر بیشتر از فاصله‌ی بین یک ChC و PaC آن است.

در بخش پالایش محتوا پایه، یکنواختی تمام روابط IsA موجود در هستان‌شناسی به کمک اندازه‌گیری "درجه IsA" تمام یال‌ها حذف می‌شود. پس از آن، شباهت معنایی بین دو مفهوم (بر اساس وزن‌های داده‌شده) به‌منظور تعیین اقلام مشابه با پروفایل کاربر هدف، مورد ارزیابی قرار می‌گیرد. از سوی دیگر، خوشه‌بندی و هستان‌شناسی بهبودیافته، الگوریتم پیشنهادی جهت یافتن کاربران مشابه (*k-NN* بهبود یافته)، و یافتن شباهت معنایی بین دو مفهوم در گراف هستان‌شناسی، فرایندهای اصلی در بخش پالایش مشارکتی روش پیشنهادی می‌باشند. به‌منظور خوشه‌بندی کاربران از اطلاعات مربوط به رتبه‌بندی (صریح) کاربران و ویژگی‌های محتوا پایه فیلم‌ها، استفاده می‌شود. در روش خوشه‌بندیِ پیشنهادشده، هم‌پوشانی در خوشه‌بندی و دیگر اشکالات مطرح در خوشه‌بندی سنتی، حذف می‌شوند.

در مرحله بعد، هستان‌شناسی اقلام با استفاده از مرحله خوشه‌بندی، بهبود می‌یابد. در این مرحله یک ویژگی به نام خوشه‌بندی-کاربر[1] یا "UC" به هستان‌شناسی اقلام اضافه خواهد شد. این ویژگی از اقلام، شامل کاربرانی است که قلم مورد نظر را خریداری کرده یا به آن علاقه‌مند هستند. در ادامه، کاربرانی که بیشترین شباهت را به کاربر هدف دارند (k نزدیک‌ترین کاربران همسایه به کاربر هدف)، بر اساس ویژگی یادشده، شناسایی می‌شوند. در مرحله بعد،

[1] User-Clustering

[DOI: 10.61186/jsdp.20.3.197]

تعداد N قلم برتر با توجه به نیازها و علایق کاربران مشابه به کاربر هدف، مشخص می‌شوند. برخلاف الگوریتم‌های سنتی، برای تشخیص k نزدیک‌ترین کاربران همسایه به کاربر هدف، علاوه بر این‌که لازم نیست همه خوشه‌ها را برای یافتن خوشه کاربران مشابه مورد جستجو قرار دهیم، بلکه نیازی نیست تا تمام کاربران موجود در خوشه‌ی کاربران مشابه را بررسی کنیم. برای این منظور فقط آن خوشه‌هایی مورد بررسی قرار می‌گیرند که در مجموعه خوشه‌هایی با شباهت بیشتر نسبت به کاربر هدف قرار دارند. همچنین به‌منظور پیدا کردن شبیه‌ترین کاربران به کاربر هدف درون خوشه مورد نظر، فقط کاربرانی مورد بررسی قرار می‌گیرند که بر اساس ویژگی "UC" درون یک خوشه مشترک با کاربر هدف باشند.

آدرس اینترنتی[1] (URL) فیلم‌های مختلف در وب گاه IMDb (پایگاه داده فیلم بر روی شبکه اینترنت[2]) یک شاخص منحصر به فرد است که می‌تواند نمایانگر یک قلم (فیلم) به‌خصوص در سامانه باشد. با استفاده از یک نوع سرویس خزنده وب[3] و استفاده از URLهای منحصر به فرد هر فیلم، می‌توان ویژگی‌های محتوا پایه فیلم‌ها را از پایگاه داده IMDb استخراج نمود. این ویژگی‌ها به‌منظور تولید فراداده هستان‌شناسی پایه[4] در پایگاه داده ذخیره می‌شوند. در واقع، سرویس WebSPHINX صفحات وب IMDb را بر اساس ویژگی‌های مهم هر فیلم که از پیش تعیین شده‌اند، مورد تجزیه و تحلیل قرار می‌دهد.

در این مقاله، ده ویژگی مهم از اقلام (فیلم‌ها) مورد استفاده قرار می‌گیرد که عبارتند از: ژانر، بازیگران، کشور سازنده، زمان انتشار، زمان اکران، رتبه IMDb، رنگ، کارگردان، نویسنده و زبان فیلم. پس از آن، هستان‌شناسی کاربران بر اساس هستان‌شناسی اقلام و رتبه‌بندی ضمنی کاربران و همچنین بازخورد کاربران (رتبه‌بندی صریح) که از طریق Web Proxy جمع‌آوری شده است، ساخته می‌شود.

در مرحله بعد، ابتدا هستان‌شناسی مربوط به اقلام بایستی تولید شود. برای طراحی هستان‌شناسی، از درخت مفاهیم[5] (CT) استفاده می‌شود که در آن رابطه بین اقلام با رابطه IsA مشخص می‌گردد. هر گره درخت یک قلم را نشان می‌دهد و هر یال رابطه والد-فرزندی بین دو گره را نشان می‌دهد. CT یکی از ساده‌ترین مدل‌های هستان‌شناسی است که در این مقاله مورد استفاده قرار می‌گیرد. برای طبقه‌بندی اقلام در این مقاله از روش کدگذاری UNSPSC[6] استفاده می‌شود. هر قلم دارای ویژگی‌های خاص خود مانند کد محصول در UNSPSC و کد WordNet می‌باشد. علاوه بر ویژگی‌های ذکر شده، در این مقاله از ویژگی منحصر به فرد دیگری به نام "درجه IsA" برای اقلام استفاده می‌گردد. این ویژگی نشان می‌دهد که چقدر یک فرزند توسط والدش پشتیبانی می‌شود. الگوریتم زیر به منظور کشف خودکار رابطه بین دو مفهوم، مورد استفاده قرار می‌گیرد. در اینجا وزن یال‌ها به‌عنوان "درجه IsA" در نظر گرفته می‌شود.

الگوریتم پیشنهادی برای اندازه‌گیری "درجه IsA" شامل مراحل زیر است:

۱- در مرحله نخست، تمام فرزندان یک مفهومِ والد با استفاده از استاندارد UNSPSC در درخت سلسله‌مراتبی مفاهیم، مشخص می‌شوند.

۲- سپس مجموعه‌ای از اسناد مرتبط با عنوان موضوع (مفهوم والد) را می‌یابیم. برای این منظور، می‌توان از روش‌هایی که برای کشف خودکار ارتباط بین مفاهیم، کاربرد دارد استفاده کنیم [45]. "اسناد مرتبط" به معنی اسناد مربوط به مفهومِ والد است.

برای یافتن اسناد مربوط به مفهومِ والد، عبارات شکل (۴) باید در Google جستجو شوند [45,64].

در شکل (۴)، عبارت "PaC" به معنای "مفهوم والد" و عبارت "ChC" به معنی "مفهوم فرزند" است. همچنین عبارت $("ChC", "PaC")$ نشان دهنده "$ChC\ IsA\ PaC$" است. از آنجایی که در این مقاله قصد داریم یک مبنای کلی برای اسناد به‌دست آوریم، که در آن هر عبارت (فرزند) زیر مجموعه یک عبارت دیگر (والد) باشد، لذا لازم است بدون درنظرگرفتن یک عبارت (فرزند) خاص، به جستجوی عبارات مدنظر بپردازیم.

1	\$Such PaCs as ChC({, ChC}*(, |)(or|and) ChC\$|\$) Such PaCs as ChC_1, ChC_2, ChC_3,… ChC_{n-1} and ChC_n.
2	\$ChC{, ChC}*(, |)(or|and) other PaCs\$ ChC_1, ChC_2, ChC_3,… ChC_{n-1}, ChC_n, or other PaCs.
3	\$PaCs(, |)especially ChC({, ChC}*(, |)(or|and) ChC\$|\$) PaCs, especially ChC_1, ChC_2,… ChC_{n-1} and ChC_n.
4	\$PaCs(, |)including ChC({, ChC}*(, |)(or|and) ChC\$|\$) PaCs including ChC_1, ChC_2,… ChC_{n-1}, or ChC_n.

(شکل- ۴): استخراج ChC ها برای یک PaC خاص

Figure 4: Extracting ChC(s) for a PaC.

۱- برای بقیه عبارات، می‌توان تعداد معینی از صفحات یا اسناد یافت‌شده برای هر عبارت را به‌منظور یافتن مفاهیم مورد نظر، بررسی نماییم. برای مثال، می‌توان

[1] Uniform Resource Locator
[2] Internet Movie Database
[3] WebSPHINX
[4] Ontology-Based Metadata
[5] Concept Tree
[6] UNSPSC. (visited in 2016). www.unspsc.org

[DOI: 10.61186/jsdp.20.3.197]

نخستین هزار صفحه یافته‌شده برای هر عبارت را انتخاب کنیم.

۲-در این مرحله، زیرمجموعه (فرزند) یک عبارت خاص با استفاده از روابط مرحله ۲ جستجو می‌شود. برای هر عبارت (فرزند) که در اسناد یافت شده در مرحله ۳ به دست می‌آید (و همچنین در روابط مرحله ۲ صادق است)، یک امتیاز مثبت برای ارزش آن عبارت، نسبت به والدش اضافه می‌شود.

۳-در مرحله بعدی فرض می‌شود که "ChC" یک فرزند برای عبارت "PaC" است و سپس قصد داریم میزان "$ChC\ IsA\ PaC$" را اندازه‌گیری کنیم. برای این‌کار، بایستی امتیازات به‌دست‌آمده برای ChC را به مجموع امتیازات تمام فرزندان PaC تقسیم کنیم.

گفتنی است برخی از عبارات فرزند، ممکن است در روابط مرحله ۲ قرار نداشته باشند؛ بنابراین، وزن پیش فرض یال‌ها را می‌توان برابر ۱ فرض کرد و سپس امتیاز به‌دست آمده در مرحله ۵ را با آن جمع نمود تا وزن یال مورد نظر به‌دست آید.

در بعضی از اسناد ممکن است فرزندِ "عبارتِ فرزند" به جای "عبارتِ فرزند" در مرحله ۲ تعیین گردد. در این صورت یک امتیاز (با ضریب $1/k$ (مثلاً $k = 2$)) برای "عبارت فرزند اصلی" در نظر گرفته می‌شود. به‌عنوان مثال، چنان که در درخت سلسله‌مراتبی داشته باشیم: $ChC_1\ IsA\ PaC$ و $ChC_2\ IsA\ ChC_1$، و در هنگام جستجوی اسناد با عبارت: "PaC شامل ChC_2 می‌باشد" روبه‌رو شویم، این بدان معنی است که رابطه "$ChC_1\ IsA\ PaC$" صحیح است و یک امتیاز با ضریب $1/k$ برای آن در نظر گرفته می‌شود.

یکی از مهم‌ترین کارهایی که در این مقاله انجام شده است، تکمیل هستان‌شناسی مبتنی بر خوشه‌بندی کاربران، به‌منظور دست‌یابی به پالایش مشارکتی کارآمد و دقیق مبتنی بر خوشه‌بندی است. درحقیقت، یک ویژگی جدید به نام «خوشه‌بندی-کاربر» به هستان‌شناسی اقلام در این مقاله اضافه می‌شود. از این ویژگی برای هستان‌شناسی اقلام استفاده می‌شود. در این ویژگی، کاربرانی که اقلام به‌خصوصی را خریداری کرده‌اند، بر اساس خوشه‌بندی انجام شده در مرحله قبل، در یک خوشه مشترک قرار می‌گیرند. این ویژگی به ما کمک می‌کند تا کاربران مشابه را بدون نیاز به جستجو در میان همه کاربران، برای کاربر هدف بیابیم؛ بنابراین، تنها کسانی جستجو می‌شوند که در میان کاربران خوشه‌بندی شده بر اساس ویژگی «خوشه‌بندی-کاربر» (برای اقلامی که کاربر هدف، خریداری کرده است) هستند. به‌منظور تهیه لیستی از اقلام قابل پیشنهاد بر اساس پالایش مشارکتی، می‌بایست شبیه‌ترین کاربران همسایه به کاربر هدف را تعیین کنیم. برای این منظور، می‌توان از الگوریتم‌های متفاوتی استفاده نمود. الگوریتم k-NN یکی از روش‌هایی است که می‌توان از آن برای رسیدن به این هدف استفاده نماییم. البته به‌منظور بهبود و افزایش کارایی الگوریتم مذکور، در این مقاله راهکاری ارائه می‌گردد که جهت یافتن کاربران همسایه با کاربر هدف، نیازی به جستجو در میان همه کاربران نباشد. برای انجام این‌کار، تنها کاربرانی مورد بررسی قرار می‌گیرند که در میان کاربران خوشه‌بندی شده بر اساس ویژگی «خوشه‌بندی-کاربر» (برای اقلامی که کاربر هدف، خریداری کرده است) هستند. علاوه بر این، لازم نیست همه خوشه‌های کاربران مورد بررسی قرار گیرند. برای پیدا کردن کاربران همسایه با کاربر هدف، تنها خوشه‌هایی مورد بررسی قرار می‌گیرند که درون خوشه‌هایی با بیشترین شباهت به کاربر هدف، باشند. با توجه به راهکار پیشنهادی بالا، انتظار می‌رود زمان اجرا و دقت در الگوریتم مذکور بهبود یابد و همین امر موجب بهبود عملکردِ کل سامانه پیشنهادگر گردد.

۴- تجزیه و تحلیل داده‌ها (یافته‌ها)

در این بخش، قصد داریم الگوریتم پیشنهادی را با استفاده از یک مجموعه‌داده واقعی، مورد مطالعه تجربی قرار داده و با توجه به چالش‌های مطرح شده در بخش‌های گذشته، با استفاده از معیارهای مناسب به ارزیابی الگوریتم مذکور بپردازیم.

در این قسمت قصد داریم کل روش پیشنهادی را که ترکیبی از یک سامانه مبتنی بر پرکردن مقادیر گم‌شده و یک سامانه پیشنهادگر ترکیبی هستان‌شناسی پایه است، با استفاده از یک مجموعه‌داده واقعی، مورد مطالعه تجربی قرار داده و با توجه به چالش‌های مطرح‌شده در بخش‌های گذشته، با استفاده از معیارهای مناسب به ارزیابی الگوریتم یادشده پرداخته و درنهایت نتایج کلی حاصل از روش پیشنهادی را با نتایج برخی روش‌های مشابه به‌روز ارائه‌شده در این حوزه، مورد مقایسه و تحلیل قرار دهیم.

مجموعه‌داده معیار مورد استفاده در این مقاله MovieLens 1m است. که می‌توان آن را از تارنمای MovieLens بارگیری کرد. این مجموعه‌داده شامل یک‌میلیون رتبه‌بندی برای ۳۹۵۰ فیلم است که توسط ۶۰۴۰ کاربر جمع‌آوری شده‌است. هر رتبه می‌تواند دارای یک مقدار صحیح در محدوده [1-5] باشد. مقدار "۱" اختصاص داده شده توسط کاربر به یک فیلم خاص، نشان می‌دهد که آن کاربر، فیلم مورد نظر را تا حد زیادی دوست ندارد و مقدار "۵" نشان دهنده این است که کاربر، فیلم یادشده را خیلی دوست دارد. اعداد دیگر نیز بر این اساس تعریف می‌شوند.

[DOI: 10.61186/jsdp.20.3.197]

جدول ۱: ویژگی‌های مجموعه‌داده
Table.1: Dataset Characteristics

ویژگی‌ها	مقادیر
تعداد اقلام	3950
تعداد کاربران	6040
بازه رتبه‌بندی	1… 5
میانگین رتبه‌بندی مجموعه‌داده	3.58
انحراف استاندارد رتبه‌بندی مجموعه‌داده	1.17
بازه تعداد رتبه‌بندی که هر کاربر ارسال کرده است	20. . . 2314
میانگین تعداد رتبه‌بندی‌هایی که کاربران ارسال کرده‌اند	166
انحراف استاندارد تعداد رتبه‌بندی‌هایی که کاربران ارسال کرده‌اند	193
بازه تعداد رتبه‌بندی هر قلم	1. . . 3428
میانگین تعداد رتبه‌بندی هر قلم	270
انحراف استاندارد تعداد رتبه‌بندی هر قلم	384
تعداد رتبه‌بندی‌های نامعلوم (خالی)	1,000,209
پراکندگی مجموعه‌داده	$95.53\% \ (\left[1-\frac{1000209}{3950\times 6040}\right]\times 100)$

در این مجموعه‌داده ، هر کاربر حداقل ۲۰ فیلم را رتبه‌بندی کرده‌است. ما یک دهم فیلم‌ها را به‌عنوان اقلام سرد در نظر می‌گیریم، یعنی ۳۹۵ قلم. قلمی که کمتر از بیست بار رتبه‌بندی شده باشد در اینجا یک قلم سرد محسوب می‌شود. مجموعه تمام فیلم‌های سرد را ICS[1] می‌نامیم. همچنین ما یک‌دهم کاربران را به‌عنوان کاربران سرد در نظر می‌گیریم، یعنی ۶۰۴ کاربر. مجموعه تمام کاربران سرد را UCS[2] می‌نامیم. فرض بر این است که هر کاربر حداقل بیست فیلم را رتبه‌بندی کرده است. برای ارزیابی روش پیشنهادی در شرایط شروع سرد از نوع کاربر جدید، برای هر یک از آن ۶۰۴ کاربر (که به‌صورت تصادفی انتخاب شده‌اند)، یک زیرمجموعه تصادفی از اقلام رتبه‌بندی‌شده توسط آن کاربر، انتخاب شده و این اقلام از فهرست رتبه‌بندی وی حذف می‌شوند. اندازه این فهرست حذفیات به‌گونه‌ای انتخاب می‌شود که در نهایت، اندازه فهرست رتبه‌بندی آن کاربر، برابر ۱۰ شود. یک دهم مجموعه‌داده به‌عنوان مجموعه‌داده آزمایشی کلی[3] انتخاب می‌شود. ما آن را به‌عنوان زیر مجموعهٔ TN[4] نام‌گذاری می‌کنیم. گفتنی است که ICS، UCS و TN به‌صورت تصادفی سی مرتبه تولید می‌شوند و هر آزمون را سی بار مستقل تکرار کرده و میانگین نتیجه آنها را در (جدول ۱) ارائه می‌کنیم.

[1] Item Cold Start
[2] User Cold Start
[3] General Benchmark
[4] Testing Non-Cold Start

جهت ارزیابی سامانه پیشنهادگر پیشنهادی، میانگین خطای مطلق[5] (MAE) به‌عنوان معیار ارزیابی استفاده می‌شود. معیار MAE توسط رابطه زیر محاسبه می‌شود. در رابطه بالا، R_{ij} بیانگر مقدار امتیازی است که کاربر i–ام به فیلم j–ام اختصاص می‌دهد، $\hat{R}_{ij}$ نیز معرف مقدار پیش‌بینی شده R_{ij} است، و $TestSize$ بیانگر تعداد مقادیر رتبه‌بندیِ پرشده است. یکی دیگر از معیارهای مورد استفاده جهت تحلیل روش پیشنهادی، تصحیح رتبه‌بندی[6] (RC) است، که با استفاده از رابطه زیر محاسبه می‌گردد.جدول ۱: ویژگی‌های مجموعه‌داده

$$RC=\frac{\sum_i\sum_j \delta\left(R_{ij}==round\left(\hat{R}_{ij}\right)\right)}{TestSize} \qquad (۴۰)$$

در رابطه بالا نیز $TestSize$ تعداد اقلامی است که کاربران آزمایشی به آنها امتیاز داده‌اند، R_{ij} بیانگر مقدار امتیازی است که کاربر i–ام به قلم j–ام اختصاص می‌دهد، و $\hat{R}_{ij}$ امتیاز پیش‌بینی‌شده‌ای است که کاربر i–ام به قلم j–ام می‌دهد. همچنین در رابطه بالا، اگر A درست باشد، $\delta(A)$ معادل یک و در غیر این صورت، مقدار آن صفر است. گفتنی است در انتهای این بخش یک آزمون آماری نیز برای تمامی نتایجِ به‌دست‌آمده انجام می‌شود، تا تأیید شود که نتایج حاصله، تصادفی به‌دست نیامده‌اند.

[5] Mean Absolute Error
[6] Rating Correction

[DOI: 10.61186/jsdp.20.3.197]

با توجه به نیازِ دقتِ اندازه‌گیری‌ها، معیارهای پشتیبانی تصمیم[1] می‌توانند نقش ویژه‌ای در زمینه ارزیابی سامانه‌های پیشنهادگر چندمعیاره داشته باشند. در این زمینه می‌توان از معیارهای رایج مورد استفاده جهت بازیابی داده‌ها، بهره برد. رابطه زیر، که به‌عنوان رابطه دقت شناخته می‌شود، سهم اقلام مرتبط با نتایج به‌دست‌آمده را تعیین می‌کند. همچنین رابطه فراخوانی، سهم اقلام بازیابی‌شده مرتبط را مشخص می‌کند. گفتنی است با توجه به این‌که با افزایش تعداد اقلام بازیابی‌شده، فراخوانی افزایش یافته و دقت کاهش می‌یابد، لذا می‌بایست از معیاری استفاده شود که قادر باشد هر دو معیار دقت و فراخوانی را به‌صورت مشترک با هم در نظر بگیرد:

$$Pre = \frac{TPR}{TPR + FPR} \quad (۴۱)$$

$$Rec = \frac{TPR}{TPR + FNR} \quad (۴۲)$$

در روابط بالا FNR مقدار پیش‌بینی‌های اشتباه غیرمرتبط است، TPR مقدار پیش‌بینی‌های درست مرتبط، و FPR مقدار پیش‌بینی‌های اشتباه مرتبط است.

معیار فیشر[2] هر دو مقدار Pre و Rec را با هم در نظر می‌گیرد [65] و طبق رابطه زیر محاسبه می‌گردد؛ در واقع معیار مذکور بیانگر میانگین هارمونیک معیارهای دقت و فراخوانی است. پارامتر Φ می‌تواند برای وزن‌دادن به تأثیر یک یا هر دو این معیارها، مورد استفاده قرار گیرد. چنانچه $\Phi > 1$ باشد اهمیت معیار دقت را افزایش می‌دهد و چنانچه $\Phi < 1$ باشد تأثیر معیار فراخوانی را افزایش می‌دهد. برای دست‌یابی به یک معیار فیشر متعادل، می‌توان $\Phi = 1$ را در نظر گرفت.

$$F_{\Phi} = \frac{(1 + \Phi^2) \times Pre \times Rec}{\Phi^2 \times (Pre + Rec)} \quad (۴۳)$$

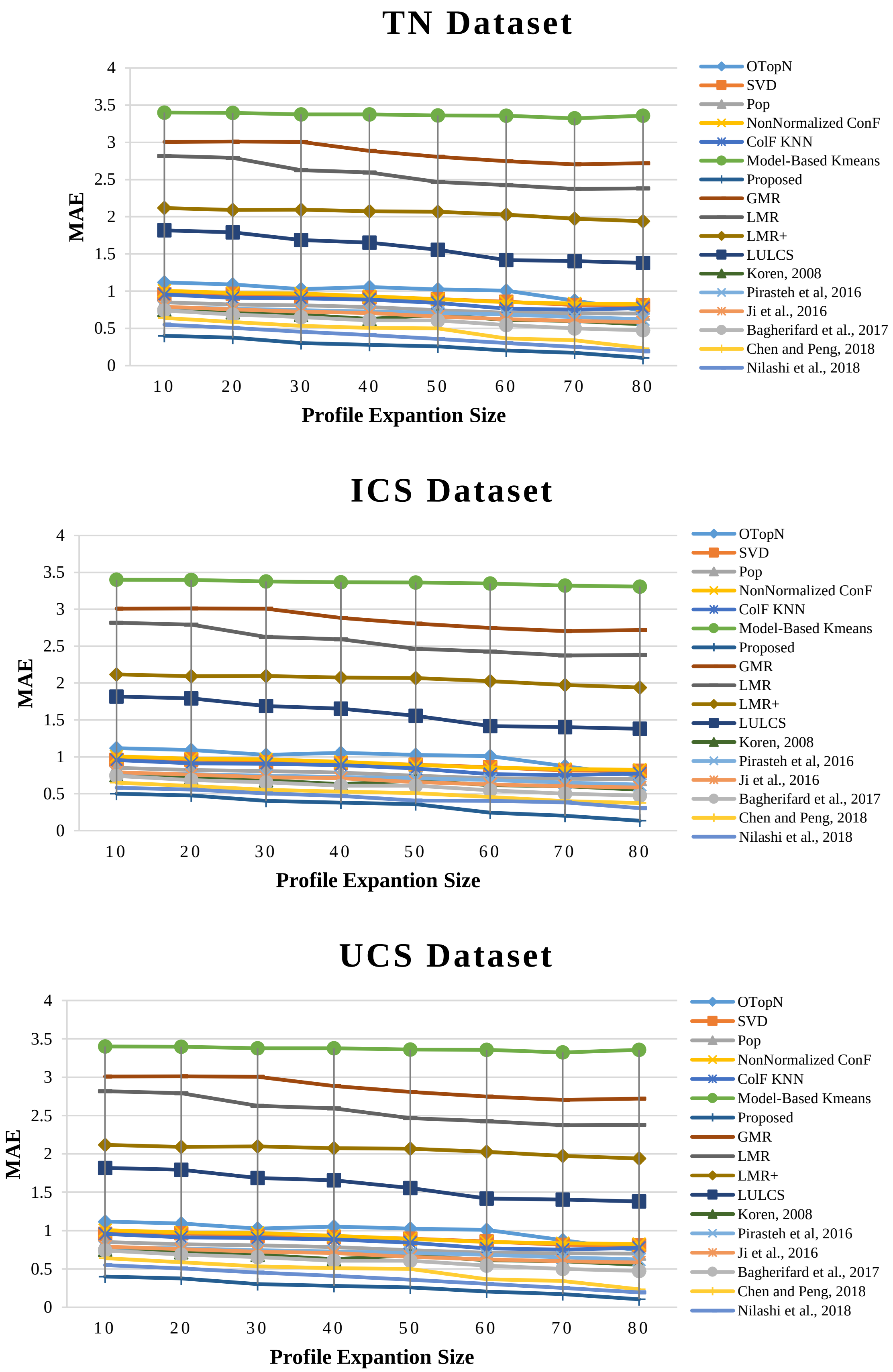
پس از مشخص‌شدن معیارهای ارزیابی، در ادامه این بخش برخی از الگوریتم‌های پایه مانند LMR، LULCS، GMR و +LMR که همان LMR غنی‌شده با اطلاعات جمعیت‌شناختی است [66-71] و برخی سامانه‌های پیشنهادگر پیچیده جدید همچون سامانه پیشنهادگر پالایش محتوا پایه غیرنرمال (NonNormalized ConF) [72]، سامانه پیشنهادگر مبتنی بر تجزیه مقدار منفرد (SVD) [73]، سامانه پیشنهادگر مبتنی بر محبوبیت (Pop) [74]، و سامانه پیشنهادگر هستان‌شناسی پایه با استفاده از فاکتورسازی ماتریس (OTopN) [75]، و همچنین روش‌های مطرح شده در شش پژوهش مشابه دیگر [76-81] را جهت مقایسه با روش پیشنهادی ارائه‌شده، انتخاب می‌کنیم. برای همه این روش‌ها، ما از بهترین پارامترهای پیشنهادشده توسط مقالات مربوطه آنها استفاده می‌کنیم. آزمایش‌ها به‌طور مستقل بر روی سه مجموعه‌داده ICS، UCS و TN انجام می‌شود. معیارهای MAE، RC، فراخوانی، معیار فیشر و زمان اجرا معیارهای ارزیابی اصلی این مقاله هستند.

شکل(۵) به منظور مقایسه عملکرد سامانه‌های پیشنهادگر مختلف، مقدار معیار MAE را برای اندازه‌های مختلف گسترش پروفایل کاربر نمایش می‌دهد. همان‌طور که مشاهده می‌شود، برای اندازه گسترش پروفایل ۱۰ و روش پیشنهادی در مجموعه‌داده UCS، مقدار معیار MAE حدود ۰.۴ است و با افزایش اندازه گسترش پروفایل به ۲۰ مقدار معیار MAE کاهش‌یافته و به حدود ۰.۳۸ می‌رسد، به همین نحو مقدار MAE با افزایش اندازۀ گسترش پروفایل مدام در حال کاهش است؛ همه روش‌های دیگر نیز به همین نحو با افزایش سایز گسترش پروفایل در هر سه مجموعه‌داده ، مقدار MAE آنها کاهش می‌یابد. به‌طور کلی همان‌طور که مشاهده می‌شود مقدار معیار MAE روش پیشنهادی در مجموعه‌داده ICS کمی بیشتر از مجموعه‌داده TN است، که علت این امر وجود داده‌های سرد در مجموعه‌داده ICS است؛ اما در کل، مقدار معیار MAE روش پیشنهادی نسبت به سایر روش‌ها، مقدار کمتری دارد و این بدان معنی است که روش پیشنهادی، دقت پیش‌بینی رتبه‌بندی بهتری نسبت به سایر روش‌های مشابه دارد.

در (شکل۶) نیز به منظور مقایسه عملکرد سامانه‌های پیشنهادگر مختلف، مقدار معیار RC برای گسترش پروفایل کاربر با اندازه ۱۰ در سه مجموعه‌داده TN، ICS و UCS نمایش داده شده‌است. در این نمودار سه روش برتر در هر سه مجموعه‌داده TN ، ICS و UCS به ترتیب: روش پیشنهادی، روش نیلاشی و همکاران، و روش چن و پنگ هستند. همانطور که می‌بینید این سه روش در همه مجموعه‌داده ها بهتر عمل کرده‌اند، علت برتری دو روش اول استفاده از هستان‌شناسی است و علت حضور روش چن و پنگ، در این سه روش برتر استفاده از بازخوردهای ضمنی و صریح بوده‌است. از طرف دیگر همان‌طورکه می‌بینید روش‌های گوناگون کارایی بالاتری بر روی مجموعه‌داده TN دارند، سپس کارایی آنها بر روی مجموعه‌داده ICS بهتر هستند و در نهایت بدترین نتایج بر روی مجموعه‌داده UCS به‌دست آمده است؛ علت این امر وجود داده‌های سرد در مجموعه‌داده ICS و UCS است و همچنین علت برتری نتایج روش‌های گوناگون در مجموعه‌داده ICS نسبت به UCS وجود توصیفات اقلام است.

[1] Decision Support
[2] F-Measure

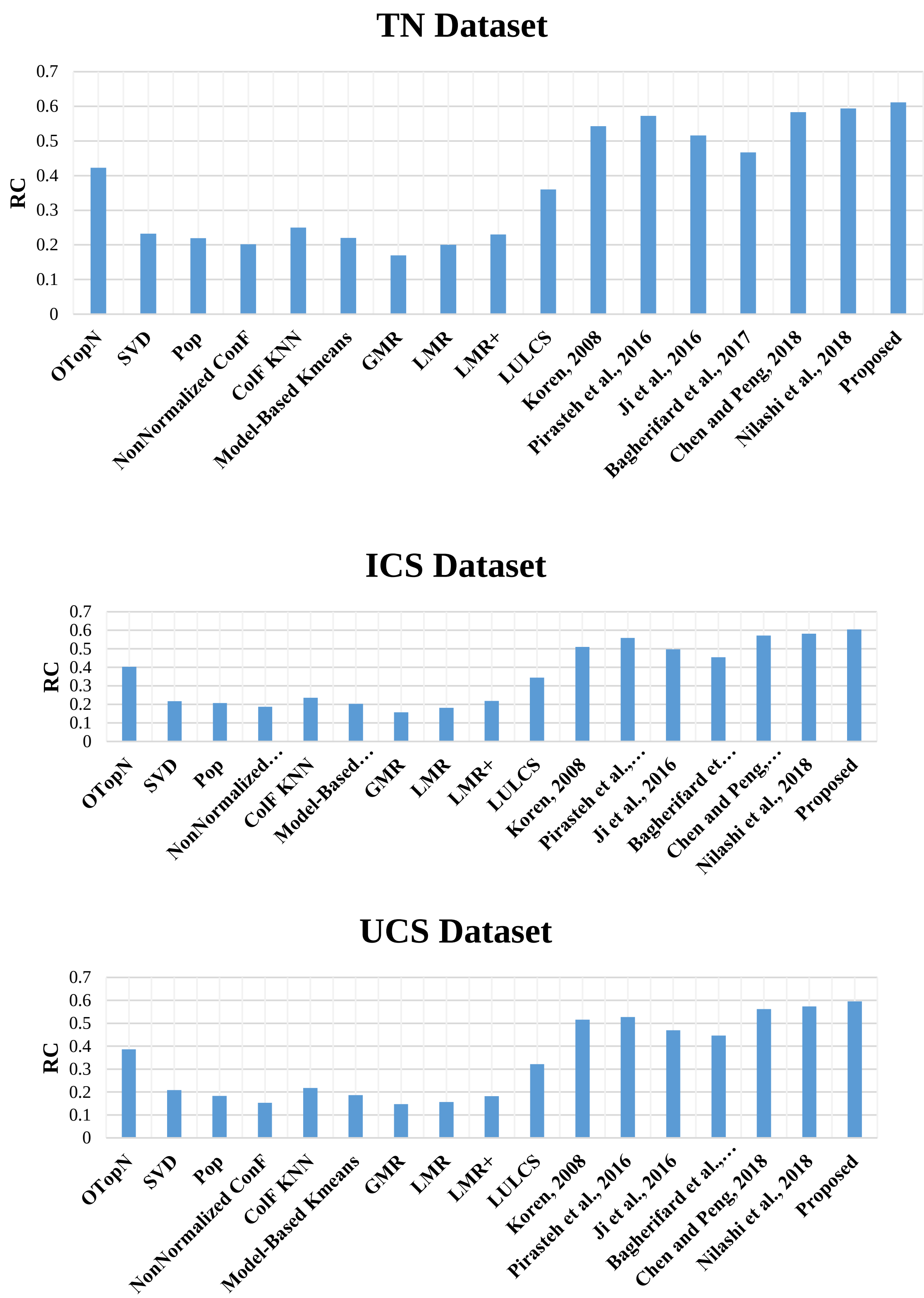
[DOI: 10.61186/jsdp.20.3.197]



(شکل – ۵): مقایسه عملکرد سامانه‌های پیشنهادگر مختلف بر حسب معیار MAE برای اندازه‌های مختلف گسترش پروفایل کاربر در (بالا) مجموعه‌داده TN، (وسط) مجموعه‌داده ICS و (پایین) مجموعه‌داده UCS

Figure 5: The performance comparison of different RSs in terms of MAE for different profile expansion sizes on a) (top) TN dataset, b) (middle) ICS dataset, and c) (bottom) UCS dataset.

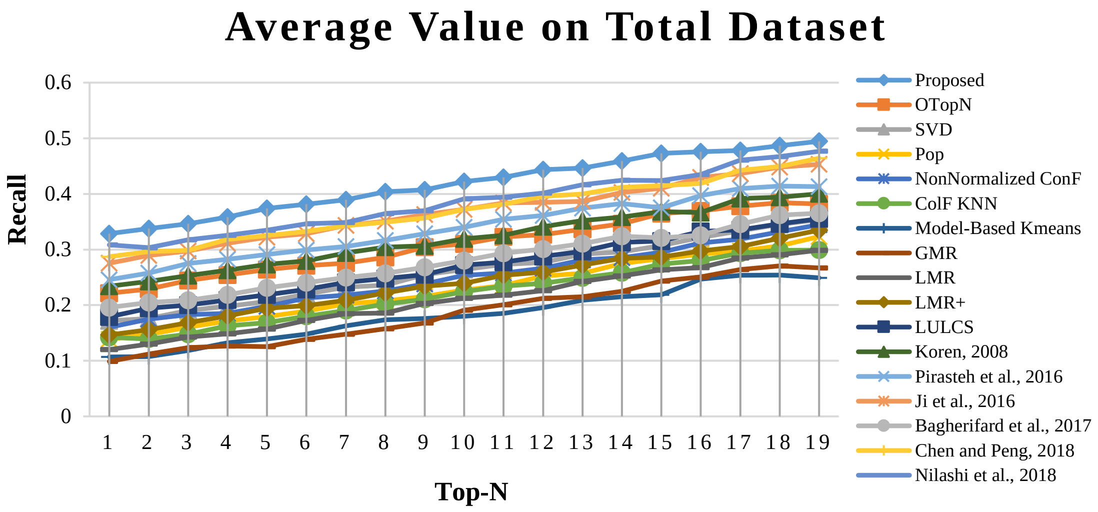
[DOI: 10.61186/jsdp.20.3.197]



(شکل-۶): مقایسه عملکرد سامانه‌های پیشنهادگر مختلف بر حسب معیار RC برای گسترش پروفایل کاربر با اندازه ۱۰ در (بالا) مجموعه‌داده TN، (وسط) مجموعه‌داده ICS و (پایین) مجموعه‌داده UCS

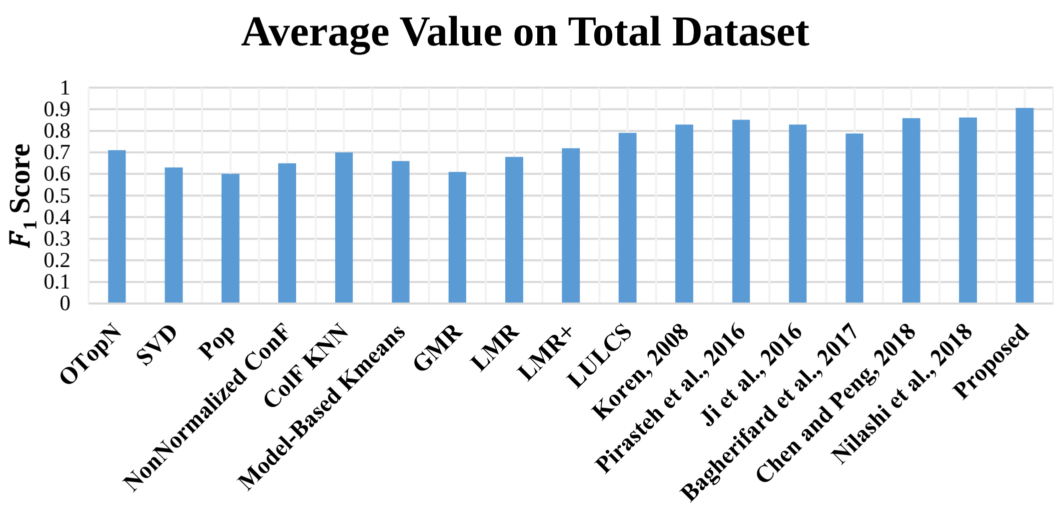
Figure 6: The performance comparison of different RSs in terms of RC for profile expansion size of 10 on a) (top) TN dataset, b) (middle) ICS dataset, and c) (bottom) UCS dataset

[DOI: 10.61186/jsdp.20.3.197]



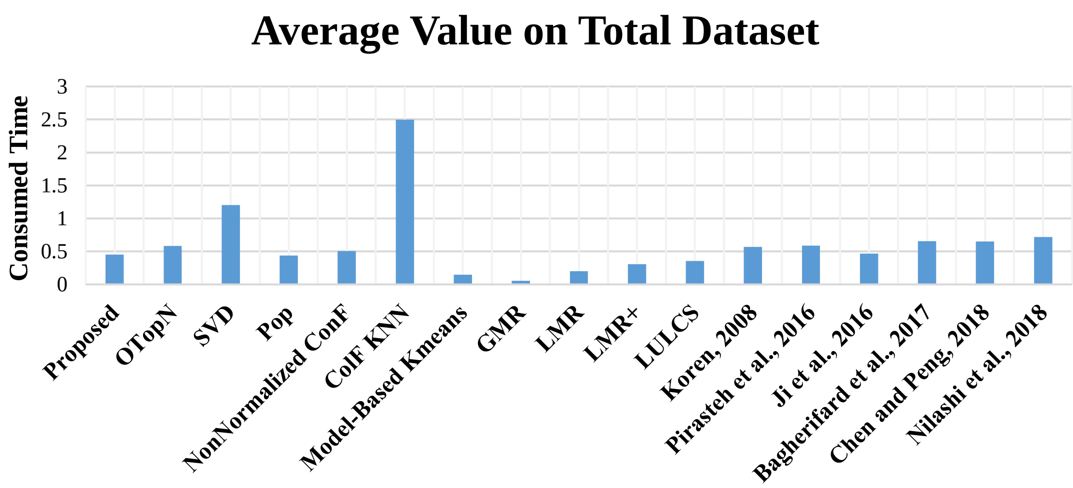
(شکل – ۷): مقایسه عملکرد سامانه‌های پیشنهادگر مختلف بر حسب معیار فراخوانی برای اندازه‌های مختلف تعداد N پیشنهاد برتر، برای گسترش پروفایل کاربر با اندازه ۱۰ برحسب میانگین در تمام مجموعه‌داده ها

Figure 7: The performance comparison of different RSs in terms of Recall of Top-N recommendations for profile expansion size of 10 on Average Value on Total dataset



(شکل–۸): مقایسه عملکرد سامانه‌های پیشنهادگر مختلف بر حسب میانگین معیار F1 در تمام مجموعه‌داده ها

Figure 8: The performance comparison of different RSs in terms of F1-measure averaged across all datasets



(شکل – ۹): مقایسه عملکرد سامانه‌های پیشنهادگر مختلف بر حسب زمان اجرا

Figure 9: The performance comparison of different RSs in terms of Consumed Time

[DOI: 10.61186/jsdp.20.3.197]

شکل (۷) به معیار فراخوانی Top-N بر حسب میانگین در تمام مجموعه‌داده‌ها اختصاص دارد. همان‌طورکه ملاحظه می‌کنید در اینجا نیز سه روش برتر به‌ترتیب: روش پیشنهادی، روش نیلاشی و همکاران، و روش چن و پنگ هستند و سه روش بدتر شامل: روش مدل پایه مبتنی بر خوشه‌بندی K-means ، GMR و LMR هستند، که این امر هم‌خوانی کامل با نتایج (شکل۵) و (شکل۶) دارد. درنهایت، میانگین معیار فیشر در تمام مجموعه‌داده‌ها برای روش‌های مختلف در شکل(۸) ارائه شده‌است. در اینجا نیز سه روش برتر به‌ترتیب: روش پیشنهادی، روش نیلاشی و همکاران، و روش چن و پنگ بوده و روش‌های بدتر شامل: Pop، GMR، SVD، روش مدل پایه مبتنی بر خوشه‌بندی K-means و روش LMR هستند، که این امر هم‌خوانی کامل با نتایج (شکل۵) و (شکل۶) دارد.

شکل (۹) زمان اجرای روش‌های مختلف را نشان می‌دهد، همان‌طورکه ملاحظه می‌کنید، سریع‌ترین روش GMR است که این امر می‌تواند به‌دلیل عدم نیاز به پردازش به‌خصوص برای این روش باشد، چون این روش همه مقادیر گمشده را با یک رویکرد سراسری مدیریت می‌کند. همچنین بدترین روش ColF KNN است که این روش اگرچه کیفیت نتایج خوبی دارد، اما بسیار کند است. در مقابل روش GMR کیفیت خروجی پایینی دارد. این امر در مورد دومین روش سریع، روش Model-Based Kmeans نیز صادق است.

آزمون فریدمن یک آزمون ناپارامتری است که برای مقایسه سه یا بیش از سه گروه وابسته که حداقل در سطح رتبه‌ای اندازه‌گیری می‌شوند، مورد استفاده قرار می‌گیرد. این آزمون می‌تواند در مورد داده‌های پیوسته (فاصله‌ای یا نسبی) نیز به‌کار برده شود، اما در هنگام محاسبه این داده‌ها نیز رتبه‌بندی آنها مدنظر قرار می‌گیرد. آزمون فریدمن معادل ناپارامتری آزمون F وابسته در تحلیل واریانس اندازه‌های تکراری است. در این حالت برای اجرای تحلیل واریانس داده‌های تکرار شده ضرورتی به‌وجود فرضیاتی مانند نرمال بودن توزیع، برابری واریانس‌ها و پیوسته بودن مقیاس وجود ندارد. بنابراین در تحلیل واریانس اندازه‌های تکراری چنانچه یک یا همه فرضیات ابتدایی مذکور رد شوند، از آزمون فریدمن استفاده می‌شود. فرضیه صفر (H0) در این آزمون بیان می‌کند که توزیع مشاهدات در سنجش‌های تکرارشده یکسان هستند. یا به‌عبارت دیگر میان توزیع‌های ایجاد شده در اثر سنجش‌های مکرر روی یک گروه و یا بین گروه‌های همتا در زمینه متغیر وابسته تفاوتی وجود ندارد. همچنین یک فرضیه جایگزین(H1) وجود دارد که با فرضیه صفر رقابت می‌کند و پیشنهاد می‌کند حداقل یکی از نتایج متفاوت است. در این مقاله، نتایج محاسبه شده با استفاده از روش پیشنهادی و سایر رویکردها، مقایسه و تحلیل می‌شود که آیا تفاوت بین نتایج با استفاده از آزمون t و فریدمن از نظر آماری معنی‌دار است؟ آزمون t یا فریدمن یک آزمون آماری است که برای ارزیابی این‌که تفاوت بین دو مجموعه‌داده تصادفی است یا از نظر آماری معنی‌دار است. محاسبه آماره فریدمن که آن را با x_r^2 نشان می‌دهند با استفاده از رابطه زیر امکان‌پذیر است:

$$x_r^2 = \frac{SS_{br}}{k(k+1)/12} \qquad (۴۴)$$

که در آن SS_{br} مجموع مجذورات رتبه‌ای بین توزیع‌ها و k تعداد توزیع‌ها می‌باشد که رتبه‌بندی در مورد آنها صورت می‌گیرد.

در ادامه نتایج آزمون آماری معناداری برای تمامی نتایجِ به‌دست آمده مربوط به روش پیشنهادی و برخی روش‌های مشابه به‌روز در (جدول۲) ارائه شده است که حاکی از آن است که نتایج حاصله، تصادفی به‌دست نیامده‌اند. در واقع این آزمون معناداری، روشی آماری برای تأیید اعتبار این است که تفاوت بین عملکرد دو یا چند روش از نظر آماری با سطح اطمینان (1-p) چقدر است. آزمون آماری معناداری را می‌توان با معیارهای مختلف ارزیابی انجام داد. اصطلاح در سطح اطمینان p به معنی با احتمال p است. (جدول۲) نتایج آزمون t را با استفاده از نتایج محاسبه شده با رویکرد پیشنهادی و سایر رویکردها ارائه می‌دهد. برای این آزمون، فرضیه صفر H0 یعنی این‌که، معنی نتایج روش‌های مختلف برابر است و H1 یعنی این‌که، معنی نتایج روش‌های مختلف نابرابر است. مقادیر $P\ (T \leq t)$ (مقدار p) بسیار نزدیک به صفر نشان می‌دهد که اطمینان ارزیابی بالاتر از ۹۹ درصد است، زیرا $P\ (T \leq t) \ll 0.05$ است. همچنین، مقادیر t-stat بزرگتر از صفر است و مقدار برگشتی $h = 1$ نشان می‌دهد که آزمون t فرضیه صفر را در سطح معنی‌داری 5% یعنی $\alpha = 0.05$ رد می‌کند. از این‌رو، رویکرد ما به‌طور قابل توجهی سایر رویکردها را بهبود بخشید. (جدول۲) نتایج به‌دست آمده از آزمون فریدمن را برای $\alpha = 0.05$ نشان می‌دهد. نتیجه توصیف می‌کند که همه رویکردها میانگین منحصر به‌فردی دارند و مقدار p بسیار نزدیک به صفر است که آزمون فرضیه صفر را رد می‌کند.

(جدول- ۲): آزمون آماری روش پیشنهادی و برخی روش‌های مشابه به‌روز

Table 2: Statistical test for the proposed method the state-of-the-art methods

Friedman's test		$t - test$					
	Nilashi et al. [81]	**Chen and Peng. [79]**	**Ji et al. [78]**	**Pirasteh et al. [76]**	**Koren. [77]**	**Bagherifard et al. [80]**	
–	0.9969	0.9960	0.9959	0.9951	0.9947	0.9901	**Mean**
1	1	1	1	1	1	1	**h**
–	3.6619	3.0530	6.3346	9.0459	12.3997	6.0955	**t - stat**
3.0774e−04	0.0026e−01	0.0037e−01	0.0032e−01	8.2749e−04	2.4317e−04	0.0037e−01	**p - value**

(جدول- ۳): مقایسه روش پیشنهادی با یکی از روش‌های مشابه به‌روز بر اساس معیار MAE

Table 3: Comparison of the proposed method with one of the updated similar methods based on the MAE criterion

Method	**Number of *k-NN***						
	5	**10**	**15**	**20**	**50**	**60**	**80**
Proposed	**0.4985320**	**0.4969234**	**0.4905253**	**0.506537**	**0.4393580**	**0.4002109**	**0.4000201**
[83]	0.5502844	0.5569947	0.5005291	0.558579	0.4583577	0.4006688	0.4000473

مقایسه روش پیشنهادی با مشابه‌ترین روش

در این بخش روش پیشنهادی با یکی از جدیدترین و مشابه‌ترین روش ارائه‌شده توسط پژوهش‌گران مقایسه شده‌است. روش ارائه‌شده در [82] یکی از سامانه پیشنهادگر حافظه پایه مبتنی بر هستان‌شناسی است. (جدول۳) روش پیشنهادی را با روش مشابه، از نظر معیار MAE مقایسه کرده‌است.

نتایج جدول(۳) نشان می‌دهند روش پیشنهادی از روش مشابه عملکرد بهتری را از خود نشان داده‌است.

۵- نتیجه‌گیری

در این مقاله، سامانه پیشنهادگر ترکیبی پالایش مشارکتی و پالایش محتوا پایه پیشنهاد شده است. به‌منظور غلبه بر مسئله شروع سرد ایده استفاده از تکنیک هستان‌شناسی در پالایش محتوا پایه مطرح شده و همچنین جهت رفع مشکلاتی همچون پراکندگی داده‌ها و سرعت اجرا در پالایش مشارکتی، از تکنیک‌های خوشه‌بندی استفاده شده‌است.

در روش پیشنهادی، روش حافظه پایه و مدل پایه با استفاده از هستان‌شناسی ترکیب شده‌است. با توجه به راه‌کارهای ارائه‌شده و نتایج به‌دست‌آمده، می‌توان گفت روش ترکیبی پیشنهادی در مقایسه با رویکردهای استاندارد حافظه پایه و مدل پایه، دارای سرعت اجرا و دقت خوبی است و بهبود قابل توجهی در مورد زمان محاسبات، دقت پیش‌بینی رتبه‌بندی، و تطابق پیشنهادهای ارائه‌شده به کاربر فعال با سلایق وی، ایجاد کرده است. بهبود پارامترهای یادشده نشان‌دهنده اثربخشی روش پیشنهادی در کاهش پراکندگی داده‌ها و افزایش مقیاس‌پذیری، در مقایسه با روش‌های مبتنی بر مدل و حافظه پایه است؛ علاوه بر راه‌کارهای ارائه‌شده در بالا، به‌منظور بهبود دقت پیشنهادها (در مقایسه با روش مدل پایه) از تکنیک‌های خوشه‌بندی استفاده شده‌است. همچنین به‌منظور ارائه سرعت اجرای قابل قبول در روش پیشنهادی (در مقایسه با روش‌های حافظه پایه) از نوعی الگوریتم حافظه پایه یعنی k-NN بهبود یافته استفاده می‌شود.

به‌اختصار می‌توان گفت در روش ترکیبی پیشنهادی، هستان‌شناسی و خوشه‌بندی پیشرفته برای تولید نتایج دقیق‌تر استفاده شده‌است. به‌منظور تجزیه و تحلیل روش پیشنهادی و کاربردی بودن آن، دقت پیش‌بینی انجام‌شده و میزان زمان اجرا، روش پیشنهادی در یک مجموعه‌داده واقعی در زمینه سامانه پیشنهادگر فیلم (ارائه‌شده توسط MovieLens) مورد ارزیابی قرار گرفت؛ همچنین به‌منظور بررسی عملکرد روش ارائه‌شده در این مقاله، با استفاده از معیارهای دقت، فراخوانی، MAE، و معیار فیشر، روش پیشنهادی با روش‌های مشابه به‌روزِ ارائه‌شده در این حوزه، مقایسه شد. با توجه به راه‌کارهای ارائه‌شده، مشخص شد روش‌های ترکیبی پالایش مشارکتی (حافظه پایه و مدل پایه) به‌همراه روش‌های محتوا پایه (بر اساس هستان‌شناسی)، دقت پیش‌بینی و زمان اجرای سامانه پیشنهادگر را بهبود می‌دهند؛ به‌منظور اثبات این قضیه با

[DOI: 10.61186/jsdp.20.3.197]

استفاده از اندازه‌گیری سه معیار: MAE، دقت تصمیم‌گیری (اندازه‌گیری پارامتر F1) و زمان اجرا، نشان داده شد روش پیشنهادی در مقایسه با روش‌های مشابه پیشین دارای عملکرد بهتری (دقت پیشنهادات ارائه شده به کاربر فعال) است. از آزمون معناداری آماری به اختلاف معنادار این روش‌ها با اطمینان ۰.۹۹ درصدی پی می‌بریم.

در آخر بیان یک نکته الزامی است، همان‌طور که در اکثر سامانه‌های پیشنهادگر، رابطه متقابل بین پارامترهای زمان محاسبات و دقت (کاهش پراکندگی داده‌ها) از اهمیت خاصی برخوردار است، ولی یک شرط حیاتی برای آن وجود دارد. با توجه به آن‌که اغلب سامانه‌های پیشنهادگر به‌عنوان بخشی از یک سامانه جامع‌تر به شکل برخط[42] مورد استفاده قرار می‌گیرند و با درنظرگرفتن این اصل که به‌طوراساسی سامانه پیشنهادگر برای بهبود تجربه کاربر در استفاده از کل سامانه به کار می‌رود، فرایند ارائه پیشنهادها بایستی در زمان منطقی انجام شود؛ بنابراین با توجه به مطالب بالا، در این مقاله نیز سعی شده‌است در کنار ارائه راه‌کارهایی جهت بهبود عملکرد سامانه پیشنهادگر، شرط حیاتی بالا را که همانا بهبود تجربه کاربر در استفاده از کل سامانه است، مد نظر قرار داده و هدف اصلی و کاربردی روش پیشنهادی خود را ارائه سامانه پیشنهادگری که بتواند به‌عنوان یک سامانه هوشمند مؤثر برای پیشنهاد اقلام (فیلم‌ها) عمل کند، قرار دهیم.

نتایج تجربی، برتری سامانه پیشنهادی را نسبت به الگوریتم‌های مشابه به‌روز از نظر دقت پیش‌بینی نشان می‌دهد و این امر را از طریق اندازه‌گیری معیارهای مختلف عملکرد و سرعت سامانه پیشنهادی را از طریق معیار زمان اجرا، بررسی می‌کند. سریع‌ترین روش مبتنی بر K-means به‌طور متوسط کمتر از ۱ (حدود ۰.۱۱) ثانیه برای اجرا مصرف می‌کند، ولی Recall آن حدود ۰.۳۰ است در حالی روش ما حدود ۰.۴ ثانیه زمان لازم دارد ولی Recall ما ۰.۸۰ است. همچنین این نتایج بیان‌گر این موضوع است که روش پیشنهادی از دقیق‌ترین روش‌ها (از جمله OtopN با Recall حدود ۰.۶۵ و زمان مصرفی حدود ۰.۵۸ ثانیه)، سریع‌تر و کیفیت آن نیز قابل رقابت و یا حتی بهتر است.

در این مقاله، فقط از روش خوشه‌بندی K-means، جهت خوشه‌بندی در سامانه پیشنهادگر استفاده شده است. ازاین‌رو در کارهای آینده می‌توان، سایر روش‌های خوشه‌بندی به‌ویژه روش‌های خوشه‌بندی جمعی را به‌منظور ارتقای سامانه پیشنهادگر، مورد توجه قرار داد.

برای این مقاله، از اطلاعات جمعیت‌شناختی[43] در سامانه پیشنهادی استفاده نشده است. بنابراین، برای تحقیقات آینده، استفاده از اطلاعات جمعیت‌شناختی کاربر می‌تواند مورد بررسی قرار گیرد تا در سامانه پیشنهادگر ترکیبی گنجانده شود؛ همچنین در این مقاله، سامانه پیشنهادی با استفاده از داده‌های مربوط به پیشنهاد فیلم، مورد ارزیابی قرار گرفته است؛ بنابراین، به‌عنوان یکی دیگر از تحقیقات آینده، می‌توان به‌منظور ارزیابی این روش، بر سایر انواع داده‌ها، مانند حوزه خرید و گردشگری تمرکز و همچنین می‌توان از سایر مدل‌های هستان‌شناسی و نیز سایر روابط در یک هستان‌شناسی به‌عنوان روش اندازه‌گیری مشابهت مفاهیم استفاده کرد.

۶- مراجع 6- Refrence

[1] E. Rich, "User modeling via stereotypes," *Cognitive science*, vol. 3, no. 4, pp. 329-354, 1979.

[2] F. McSherry and I. Mironov, "Differentially private recommender systems: Building privacy into the netflix prize contenders," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009: ACM, pp. 627-636.

[3] S. S. Anand and B. Mobasher, "Intelligent techniques for web personalization," in *Proceedings of the 2003 international conference on Intelligent Techniques for Web Personalization*, 2003: Springer-Verlag, pp. 1-36.

[4] C.-P. Wei, C.-S. Yang, and H.-W. Hsiao, "A collaborative filtering-based approach to personalized document clustering," *Decision Support Systems*, vol. 45, no. 3, pp. 413-428, 2008.

[5] M. López-Nores *et al.*, "MiSPOT: dynamic product placement for digital TV through MPEG-4 processing and semantic reasoning," *Knowledge and Information Systems*, vol. 22, no. 1, pp. 101-128, 2010.

[6] L. Liu, N. Mehandjiev, and D.-L. Xu, "Multi-criteria service recommendation based on user criteria preferences," in *Proceedings of the fifth ACM conference on Recommender systems*, 2011: ACM, pp. 77-84.

[7] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative filtering recommender systems," in *The adaptive web*: Springer, 2007, pp. 291-324.

[8] L.-C. Cheng and H.-A. Wang, "A fuzzy recommender system based on the integration of subjective preferences and objective information," *Applied Soft Computing*, vol. 18, pp. 290-301, 2014.

[42] Online

[43] Demographic Information

[DOI: 10.61186/jsdp.20.3.197]

market forecasting," *Applied Soft Computing*, vol. 40, pp. 132-149, 2016.

[22] S. K. Shinde and U. Kulkarni, "Hybrid personalized recommender system using centering-bunching based clustering algorithm," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1381-1387, 2012.

[23] P. Wang, "A personalized collaborative recommendation approach based on clustering of customers," *Physics Procedia*, vol. 24, pp. 812-816, 2012.

[24] N. Guarino, D. Oberle, and S. Staab, "What is an ontology?," in *Handbook on ontologies*: Springer, 2009, pp. 1-17.

[25] T. R. Gruber, "Toward principles for the design of ontologies used for knowledge sharing?," *International journal of human-computer studies*, vol. 43, no. 5-6, pp. 907-928, 1995.

[26] W. Borst, "Construction of Engineering," ed: Ontologies, University of Tweenty, Enschede, NL-Center for Telematica and Information Technology, 1997.

[27] A. Flahive, B. O. Apduhan, J. W. Rahayu, and D. Taniar, "Large scale ontology tailoring and simulation in the Semantic Grid Environment," *International Journal of Metadata, Semantics and Ontologies*, vol. 1, no. 4, pp. 265-281, 2006.

[28] G. Antoniou and F. Van Harmelen, *A semantic web primer*. MIT press, 2004.

[29] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure, "Ontological user profiling in recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 54-88, 2004.

[30] S. E. Middleton, D. De Roure, and N. R. Shadbolt, "Ontology-based recommender systems," in *Handbook on ontologies*: Springer, 2009, pp. 779-796.

[31] N. Guarino, C. Masolo, and G. Vetere, "Ontoseek: Content-based access to the web," *IEEE Intelligent Systems and their Applications*, vol. 14, no. 3, pp. 70-80, 1999.

[32] M. Craven, A. McCallum, D. PiPasquo, T. Mitchell, and D. Freitag, "Learning to extract symbolic knowledge from the World Wide Web," Carnegie-mellon univ pittsburgh pa school of computer Science, 1998.

[33] D. Godoy and A. Amandi, "User profiling for web page filtering," *IEEE Internet computing*, vol. 9, no. 4, pp. 56-64, 2005.

[34] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-based personalized search and browsing," *Web Intelligence and Agent Systems: An international Journal*, vol. 1, no. 3, 4, pp. 219-234, 2003.

[35] M. I. Martín-Vicente, A. Gil-Solla, M. Ramos-Cabrer, J. J. Pazos-Arias, Y. Blanco-Fernández, and M. López-Nores, "A semantic approach to improve neighborhood formation in collaborative recommender systems,"

[9] M. Nilashi, O. bin Ibrahim, and N. Ithnin, "Hybrid recommendation approaches for multi-criteria collaborative filtering," *Expert Systems with Applications*, vol. 41, no. 8, pp. 3879-3900, 2014.

[10] S. S. Anand, P. Kearney, and M. Shapcott, "Generating semantically enriched user profiles for web personalization," *ACM Transactions on Internet Technology (TOIT)*, vol. 7, no. 4, p. 22, 2007.

[11] C. Porcel, C. Martinez-Cruz, J. Bernabé-Moreno, Á. Tejeda-Lorente, and E. Herrera-Viedma, "Integrating ontologies and fuzzy logic to represent user-trustworthiness in recommender systems," *Procedia Computer Science*, vol. 55, pp. 603-612, 2015.

[12] P. Giaretta and N. Guarino, "Ontologies and knowledge bases towards a terminological clarification," *Towards very large knowledge bases: knowledge building & knowledge sharing*, vol. 25, no. 32, pp. 307-317, 1995.

[13] J. S. Breese, 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: Morgan Kaufmann Publishers Inc., pp. 43-52.

[14] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, vol. 2009, p. 4, 2009.

[15] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering," in *Proceedings of the fifth international conference on computer and information technology*, 2002, vol. 1, pp. 291-324.

[16] T. Hofmann and J. C. Puzicha, "System and method for personalized search, information filtering, and for generating recommendations utilizing statistical latent class models," ed: Google Patents, 2004.

[17] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.

[18] D. Zhou *et al.*, "Learning multiple graphs for document recommendations," in *Proceedings of the 17th international conference on World Wide Web*, 2008: ACM, pp. 141-150.

[19] S. Gong, "A collaborative filtering recommendation algorithm based on user clustering and item clustering," *JSW*, vol. 5, no. 7, pp. 745-752, 2010.

[20] Y. He, S. Yang, and C. Jiao, "A hybrid collaborative filtering recommendation algorithm for solving the data sparsity," in *Computer Science and Society (ISCCS), 2011 International Symposium on*, 2011: IEEE, pp. 118-121.

[21] H. J. Sadaei, R. Enayatifar, M. H. Lee, and M. Mahmud, "A hybrid model based on differential fuzzy logic relationships and imperialist competitive algorithm for stock

[DOI: 10.61186/jsdp.20.3.197]

Expert Systems with Applications, vol. 41, no. 17, pp. 7776-7788, 2014.
[36] A. Sieg, B. Mobasher, and R. Burke, "Improving the effectiveness of collaborative recommendation with ontology-based user profiles," in *proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems*, 2010: ACM, pp. 39-46.
[37] I. Cantador, A. Bellogín, and P. Castells, "A multilayer ontology-based hybrid recommendation model," *Ai Communications*, vol. 21, no. 2-3, pp. 203-210, 2008.
[38] Y. Deng, Z. Wu, C. Tang, H. Si, H. Xiong, and Z. Chen, "A hybrid movie recommender based on ontology and neural networks," in *Proceedings of the 2010 IEEE/ACM Int'l Conference on Green Computing and Communications & Int'l Conference on Cyber, Physical and Social Computing*, 2010: IEEE Computer Society, pp. 846-851.
[39] R. Burke, "Hybrid web recommender systems," in *The adaptive web*: Springer, 2007, pp. 377-408.
[40] L. Zhuhadar, O. Nasraoui, R. Wyatt, and E. Romero, "Multi-model ontology-based hybrid recommender system in e-learning domain," in *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, 2009, vol. 3: IEEE, pp. 91-95.
[41] A. Moreno, A. Valls, D. Isern, L. Marin, and J. Borràs, "Sigtur/e-destination: ontology-based personalized recommendation of tourism and leisure activities," *Engineering Applications of Artificial Intelligence*, vol. 26, no. 1, pp. 633-651, 2013.
[42] J. Lu, Q. Shambour, Y. Xu, Q. Lin, and G. Zhang, "BizSeeker: a hybrid semantic recommendation system for personalized government-to-business e-services," *Internet Research*, vol. 20, no. 3, pp. 342-365, 2010.
[43] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proceedings of the 14th international conference on World Wide Web*, 2005: ACM, pp. 22-32.
[44] T.-N. Pham, T.-H. Vuong, T.-H. Thai, M.-V. Tran, and Q.-T. Ha, "Sentiment Analysis and User Similarity for Social Recommender System: An Experimental Study," in *Information Science and Applications (ICISA) 2016*: Springer, 2016, pp. 1147-1156.
[45] P. Buitelaar, P. Cimiano, and B. Magnini, *Ontology learning from text: methods, evaluation and applications*. IOS press, 2005.
[46] D. Julong, "Introduction to grey system theory," *The Journal of grey system*, vol. 1, no. 1, pp. 1-24, 1989.
[47] R. Pan, T. Yang, J. Cao, K. Lu, and Z. Zhang, "Missing data imputation by K nearest neighbours based on grey relational structure and mutual information," *Applied Intelligence*, vol. 43, no. 3, pp. 614-632, 2015.
[48] J. M. Jou, P.-Y. Chen, and J.-M. Sun, "The gray prediction search algorithm for block motion estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 6, pp. 843-848, 1999.
[49] S.-L. Su, Y.-C. Su, and J.-F. Huang, "Grey-based power control for DS-CDMA cellular mobile systems," *IEEE Transactions on Vehicular Technology*, vol. 49, no. 6, pp. 2081-2088, 2000.
[50] Q. Song, M. Shepperd, and C. Mair, "Using grey relational analysis to predict software effort with small data sets," in *11th IEEE International Software Metrics Symposium (METRICS'05)*, 2005: IEEE, pp. 10 pp.-35.
[51] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *Journal of Systems and Software*, vol. 85, no. 11, pp. 2541-2552, 2012.
[52] C. Zhang, X. Zhu, J. Zhang, Y. Qin, and S. Zhang, "GBKII: An imputation method for missing values," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2007: Springer, pp. 1080-1087.
[53] J. K. Dixon, "Pattern recognition with partly missing data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 10, pp. 617-621, 1979.
[54] D. W. Aha and R. L. Goldstone, "Concept learning and flexible weighting," in *Proceedings of the fourteenth annual conference of the Cognitive Science Society*, 1992: Citeseer, pp. 534-539.
[55] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30-37, 2009.
[56] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear algebra*: Springer, 1971, pp. 134-151.
[57] X. Su and T. M. Khoshgoftaar, "Collaborative filtering for multi-class data using belief nets algorithms," in *2006 18th IEEE international conference on Tools with Artificial Intelligence (ICTAI'06)*, 2006: IEEE, pp. 497-504.
[58] G. Shani, D. Heckerman, R. I. Brafman, and C. Boutilier, "An MDP-based recommender system," *Journal of Machine Learning Research*, vol. 6, no. 9, 2005.
[59] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50-57.
[60] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391-407, 1990.

[DOI: 10.61186/jsdp.20.3.197]

[DOI: 10.61186/jsdp.20.3.197]

[61] L. M. Chan, S. S. Intner, and J. Weihs, *Guide to the Library of Congress classification*. ABC-CLIO, 2016.

[62] C. J. Becerra, S. Jimenez, and A. F. Gelbukh, "Towards User Profile-based Interfaces for Exploration of Large Collections of Items," *Decisions@ RecSys*, vol. 13, pp. 9-16, 2013.

[63] S. Jimenez, F. A. Gonzalez, and A. Gelbukh, "Mathematical properties of soft cardinality: Enhancing Jaccard, Dice and cosine similarity measures with element-wise distance," *Information Sciences*, vol. 367, pp. 373-389, 2016.

[64] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th conference on Computational linguistics-Volume 2*, 1992: Association for Computational Linguistics, pp. 539-545.

[65] C.-F. Tsai and C. Hung, "Cluster ensembles in collaborative filtering recommendation," *Applied Soft Computing*, vol. 12, no. 4, pp. 1417-1425, 2012.

[66] L. H. Son, "Dealing with the new user cold-start problem in recommender systems: A comparative review," *Information Systems*, vol. 58, pp. 87-104, 2016.

[67] V. Vanitha and P. Krishnan, "A modified ant colony algorithm for personalized learning path construction," *Journal of Intelligent & Fuzzy Systems*, vol. 37, no. 5, pp. 6785-6800, 2019.

[68] N. Silva, D. Carvalho, A. C. Pereira, F. Mourão, and L. Rocha, "The pure cold-start problem: A deep study about how to conquer first-time users in recommendations domains," *Information Systems*, vol. 80, pp. 1-12, 2019.

[69] J. R. Almeida, E. Monteiro, L. B. Silva, A. P. Sierra, and J. L. Oliveira, "A recommender system to help discovering cohorts in rare diseases," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2020: IEEE, pp. 25-28.

[70] J. Zhong, H. Xie, and F. L. Wang, "The research trends in recommender systems for e-learning: A systematic review of SSCI journal articles from 2014 to 2018," *Asian Association of Open Universities Journal*, 2019.

[71] L. Romero, C. Saucedo, M. Caliusco, and M. Gutiérrez, "Supporting self-regulated learning and personalization using ePortfolios: a semantic approach based on learning paths," *International Journal of Educational Technology in Higher Education*, vol. 16, no. 1, pp. 1-16, 2019.

[72] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of recommendation algorithms for e-commerce," in *Proceedings of the 2nd ACM Conference on Electronic Commerce*, 2000, pp. 158-167.

[73] P. Cremonesi, Y. Koren, and R. Turrin, "Performance of recommender algorithms on top-n recommendation tasks," in *Proceedings of the fourth ACM conference on Recommender systems*, 2010, pp. 39-46.

[74] R. Bambini, P. Cremonesi, and R. Turrin, "A recommender system for an iptv service provider: a real large-scale production environment," in *Recommender systems handbook*: Springer, 2011, pp. 299-331.

[75] H. Cui, M. Zhu, and S. Yao, "Ontology-based Top-N Recommendations on New Items with Matrix Factorization," *J. Softw.*, vol. 9, no. 8, pp. 2026-2032, 2014.

[76] P. Pirasteh, D. Hwang, and J. J. Jung, "Exploiting matrix factorization to asymmetric user similarities in recommendation systems," *Knowledge-Based Systems*, vol. 83, pp. 51-57, 2016.

[77] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 426-434.

[78] K. Ji, R. Sun, X. Li, and W. Shu, "Improving matrix approximation for recommendation via a clustering-based reconstructive method," *Neurocomputing*, vol. 173, pp. 912-920, 2016.

[79] S. Chen and Y. Peng, "Matrix factorization for recommendation with explicit and implicit feedback," *Knowledge-Based Systems*, vol. 158, pp. 109-117, 2018.

[80] K. Bagherifard, M. Rahmani, M. Nilashi, and V. Rafe, "Performance improvement for recommender systems using ontology," *Telematics and Informatics*, vol. 34, no. 8, pp. 1772-1792, 2017.

[81] M. Nilashi, O. Ibrahim, and K. Bagherifard, "A recommender system based on collaborative filtering using ontology and dimensionality reduction techniques," *Expert Systems with Applications*, vol. 92, pp. 507-520, 2018.

[82] P. Lops, M. d. Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," *Recommender systems handbook*, pp. 73-105, 2011.

[۸۳] بحرانی، پیام، مینایی بیدگلی، بهروز، پروین، حمید، میرزارضایی، میترا، و کشاورز، احمد. (۱۴۰۰). ارائه یک سامانه پیشنهادگر حافظه پایه ترکیبی با استفاده از هستان شناسی و محتوا. پردازش علائم و داده ها، ۱۸(۴ (۵۰ پیاپی))، ۸۹-۱۲۴.

ضمائم

در شکل های زیر شبه کدهای چارچوب پیشنهادی ارائه شده است.

Algorithm 1: ProposedClustering

Input:
R: An incomplete rating dataset
N: Number of clusters
Output:
R: Completed rating dataset
ρ: A partition of users

$[\alpha,\beta] = size(R)$
$\forall i \in \{1,\dots,\alpha\}, j \in \{1,\dots,\beta\}: \dot{R}_{ij} = \frac{R_{ij} - \min_{k\in\{1,\dots,\beta\}} R_{ik}}{\max_{k\in\{1,\dots,\beta\}} R_{ik} - \min_{k\in\{1,\dots,\beta\}} R_{ik}}$
$\forall i \in \{1,\dots,\alpha\}, j \in \{1,\dots,N\}: U_{ij} = random[1,\beta]$
$\forall i \in \{1,\dots,\alpha\}, j \in \{1,\dots,N\}: \dot{U}_{ij} = \frac{U_{ij}}{\sum_{k=1}^{N} U_{ik}}$
$\forall i \in \{1,\dots,\alpha\}, j \in \{1,\dots,\gamma\}: m_{ij}^{new} = 0 \,\&\, m_{ij} = \dot{R}_{ij}$
$\dot{R}_{new} = \dot{R}$
For $ite = 1$ to $MaxIteration$
    $m_{IJ}^{new} = \frac{\sum_{i=1}^{\alpha} u_{iI} \times \Phi_{iIJ}(R) \times \dot{R}_{iJ}}{\sum_{i=1}^{\alpha} u_{iI}}$
    $If\ (m == m^{new})$
        Break;
    $m = m_{IJ}^{new}$
    $\forall i \in \{1,\dots,\alpha\}, j \in \{1,\dots,N\}: U_{ij} =$
$\begin{cases} 1 & \forall k \in \{1,\dots,N\}: \mathcal{D}_{ij}(R) \le \mathcal{D}_{ik}(R) \\ 0 & O.W. \end{cases}$
    $\forall j \in \{1,\dots,N\}: \rho_j = \{\ \ \}$
    $\forall i \in \{1,\dots,\alpha\}: \lambda_i = \arg\max_{j\in\{1,\dots,N\}} U_{ij} \rightarrow \rho_{\lambda_i} = \rho_{\lambda_i} \cup \{i\}$
    For $i = 1$ to α
        $T = \rho_{\lambda_i} - \{i\}$
        For $j = 1$ to β
            $If\ (\dot{R}_{ij} = NaN)$
                $S1 = 0; S2 = 0;$
                For $q \in T$
                    $If\ (\dot{R}_{qj} \ne NaN)$
                        $\mathscr{w}_q = \frac{e^{\mathscr{b}\mathscr{s}_{iq}(R)}}{e^{\mathscr{b}}}$
                        $S1 = S1 + \mathscr{w}_q; S2 = S2 + \mathscr{w}_q \times \dot{R}_{qj};$
                $\dot{R}_{new_{ij}} = \frac{S2}{S1}$
$\forall i \in \{1,\dots,\alpha\}, j \in \{1,\dots,\beta\}: T_{ij} = \left[\max_{k\in\{1,\dots,\beta\}} R_{ik} - \min_{k\in\{1,\dots,\beta\}} R_{ik}\right] \times \dot{R}_{new_{ij}} + \min_{k\in\{1,\dots,\beta\}} R_{ik}$
$R = T$
$Return\ R, \rho$

شکل پ-۱: الگوریتم اول مربوط به شبه کد خوشه‌بندی روش پیشنهادی

Figure P-1: First algorithm related to clustering pseudocode of the proposed method

Algorithm 2: OBRS

Input:
R: An incomplete rating dataset
Des^I: Description text of different items
p: Number of keywords
τ: Algorithm parameter
n_1: Algorithm parameter
Output:
IP: The items' profiles
UP: The users' profiles
R: Completed rating dataset

$[\alpha,\beta] = size(R)$
$Compute\ [b, \forall j \in \{1,2,\dots,\beta\}: b_{-j}, \forall i \in \{1,2,\dots,\alpha\}: b_{i-}]$
$WholeDes = \cup_{i=1}^{\beta} Des_i^I$
$O = ExtractOntology(WholeDes)$
$KW = ExtractKeyWords(WholeDes, p)$
$\forall i \in \{1,2,\dots,\beta\}, j \in \{1,2,\dots,p\}: IP_{ij} = tf_{ji} \times idf_j$ according to Des_i^I and $WholeDes$
$\forall i \in \{1,2,\dots,\beta\}, j \in \{1,2,\dots,\beta\}: {}_I\dot{s}_{ij} = {}_I s_{ij}^{TFIDF}$
$\forall i \in \{1,2,\dots,\beta\}, j \in \{1,2,\dots,\beta\}: {}_I\ddot{s}_{ij} = {}_I^O s_{ij}^{M}(\tau)$
$\forall i \in \{1,2,\dots,\beta\}, j \in \{1,2,\dots,\beta\}: {}_I s_{ij} = {}_I\ddot{s}_{ij}$ and ${}_I s_{i(\beta+j)} = {}_I\dot{s}_{ij}$
$[{}_I List, {}_I Centers] = Clustering({}_I s, n_1)$
$\forall i \in \{1,2,\dots,\alpha\}: IUD_i = \{j \in \{1,2,\dots,\beta\} | R_{ij} \ge b + b_{i-} + b_{-j}\}$
$\forall j \in \{1,2,\dots,\alpha\}: Des_j^U = \cup_{i\in IUD_j} Des_i^I$
$\forall i \in \{1,2,\dots,\alpha\}, j \in \{1,2,\dots,p\}: UP_{ij} = tf_{ji} \times idf_j$ according to Des_i^U and $WholeDes$
$\forall i \in \{1,2,\dots,\alpha\}, j \in \{1,2,\dots,\beta\}: {}_I\dot{s}_{ij} = {}_{UI}s_{ij}^{TFIDF}$ *where* ${}_{UI}s_{ij}^{TFIDF}$ *=cosine similarity between* UP_i *and* IP_j
$\forall i \in \{1,2,\dots,\alpha\}, j \in \{1,2,\dots,\beta\}: {}_I\ddot{s}_{ij} = {}_{UI}^O s_{ij}^{M}(\tau)$ *where* ${}_{UI}^O s_{ij}^{M}(\tau)$ *=ontology similarity between* Des_i^U *and* Des_j^I
$\forall i \in \{1,2,\dots,\alpha\}, j \in \{1,2,\dots,\beta\}: {}_{UI}s_{ij} = {}_{UI}\ddot{s}_{ij}$ and ${}_{UI}s_{i(\beta+j)} = {}_{UI}\dot{s}_{ij}$
$\forall i \in \{1,2,\dots,\alpha\}, j \in \{1,2,\dots,\beta\}: {}_U s_{ij} = {}_{UI}\ddot{s}_{ij}$ and ${}_{UI}s_{i(\beta+j)} = {}_{UI}\dot{s}_{ij}$
$\forall i \in \{1,2,\dots,\alpha\}$: *Assign the* i*th user to an item cluster according to generalizing* ${}_{UI}s_{i:}$ *to* ${}_I Centers$
$\forall i \in \{1,2,\dots,\beta\}$: *Put indices of all users that are in a shared cluster with the* i*th item in* ${}_I UC_i$
$\forall i \in \{1,2,\dots,\alpha\}, j \in \{1,2,\dots,\alpha\}: {}_U s_{ij} = Cor(R_{i:}, R_{j:})$
$\forall i \in \{1,2,\dots,\beta\}, j \in \{1,2,\dots,\beta\}: {}_I s_{ij} = Cor(R_{:i}, R_{:j})$
$\forall i \in \{1,2,\dots,\alpha\}, j \in \{1,2,\dots,\beta\}: \hat{R}_{ij} =$
$$\begin{cases} b_{ij} + \frac{\sum_{i'\in {}_I UC_j} \pi(Cor(R_{i:},R_{i':}),0)(R_{i'j} - b_{i'j})}{\sum_{i'\in {}_I UC_j} \pi(Cor(R_{i:},R_{i':}),0)} & if\ user\ ith\ is\ not\ cold \\ b_{ij} + \frac{\sum_{j'\in NNI_j^{ik}} \pi(Cor(R_{:j},R_{:j'}),0)(R_{ij'} - b_{ij'})}{\sum_{j'\in NNI_j^{ik}} \pi(Cor(R_{:j},R_{:j'}),0)} & elseif\ jth\ item\ is\ not\ col \\ b_{ij} + \frac{\sum_{j'\in NNI_j^{ik}} \pi({}_I\dot{s}_{jj'},0)(R_{ij'} - b_{ij'})}{\sum_{j'\in NNI_j^{ik}} \pi({}_I\dot{s}_{jj'},0)} & else \end{cases}$$
$Return\ IP, UP, \hat{R}$

شکل پ-۲: الگوریتم دوم مربوط به شبه‌کد سامانه پیشنهادگر مبتنی بر آنتولوژی پیشنهادی

Figure P-2: Second algorithm related to pseudo-code of the proposed ontology-based recommender system

Algorithm 3: ProposedRS

Input:
R: An incomplete rating dataset
t: Index of target user
N_1: Number of defined clusters of users
N_2: Number of used clusters of users
N_3: Number of used similar items
Des^I: Description text of different items
p: Number of keywords
n_1: Algorithm parameter
Output:
UR: Completed rating of the target user

$[\alpha,\beta] = size(R)$
$[IP, UP, \hat{R}] = \mathbf{OBRS}(R, Des^I, p, 2, n_1)$
$\tilde{R} = \boldsymbol{MKC}(R, N)$
$UR = \frac{\hat{R}_t + \tilde{R}_t}{2}$
$\boldsymbol{Return\ UR}$

شکل پ-۳: الگوریتم سوم مربوط به شبه‌کد بخش ترکیب کننده روش پیشنهادی

Figure P-3: Third algorithm related to pseudo-code of combiner part in the proposed method

[DOI: 10.61186/jsdp.20.3.197]

پیام بحرانی دانش‌آموخته دوره دکترای تخصصی دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران در رشته مهندسی کامپیوتر گرایش سامانه‌های نرم‌افزاری است. زمینه‌های پژوهشی مورد علاقه ایشان مباحثی نظیر سامانه‌های امنیت اطلاعات، هستان‌شناسی و سامانه‌های پیشنهادگر است.

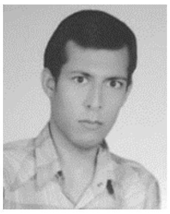
نشانی رایانامه ایشان عبارت است از:

bahranipayam@gmail.com

بهروز مینایی بیدگلی دانش‌آموخته دانشگاه ایالتی میشیگان آمریکا در رشته علوم و مهندسی کامپیوتر با تخصص هوش مصنوعی و داده‌کاوی است. وی در حال حاضر عضو هیأت علمی و دانشیار دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت و رئیس دانشکده مهندسی کامپیوتر است. ایشان سرپرستی گروه پژوهشی فناوری‌های بازی‌های رایانه‌ای و نیز آزمایشگاه داده‌کاوی را به عهده دارد. محاسبات نرم، یادگیری ماشین، بازی‌های رایانه‌ای، داده‌کاوی، متن‌کاوی، و پردازش زبان طبیعی، زمینه‌های پژوهشی مورد علاقه ایشان است.

نشانی رایانامه ایشان عبارت است از:

b_minaei@iust.ac.ir



حمید پروین تحصیلات خود را در مقطع کارشناسی در دانشگاه چمران اهواز به پایان رساند. ایشان مدرک کارشناسی ارشد و دکترا را در دانشگاه علم و صنعت دریافت کردند و پس از آن به عضویت هیأت علمی دانشگاه آزاد اسلامی واحد نورآباد ممسنی درآمدند. وی هم‌اکنون در چندین واحد دانشگاهی در رشته کامپیوتر مشغول به تدریس است. زمینه‌های پژوهشی وی مباحثی نظیر الگوریتم‌های بهینه‌سازی، طبقه‌بندی و خوشه‌بندی داده‌ها است.

نشانی رایانامه ایشان عبارت است از:

parvin@iust.ac.ir

میترا میرزارضایی استادیار دانشکده فنی و مهندسی گروه کامپیوتر دانشگاه آزاد اسلامی واحد علوم و تحقیقات تهران هستند. زمینه‌های پژوهشی مورد علاقه ایشان، یادگیری ماشین و شناسایی الگوها است.

نشانی رایانامه ایشان عبارت است از:

mirzarezaee@srbiau.ac.ir

احمد کشاورز مدارک کارشناسی و کارشناسی ارشد خود را به‌ترتیب در سال‌های ۱۳۸۰ و ۱۳۸۳ از دانشگاه شیراز و تربیت مدرس در رشته مهندسی برق و مخابرات سیستم دریافت کرد. ایشان درجه دکترای خود را در سال ۱۳۸۷ از دانشگاه تربیت مدرس در رشته مخابرات سیستم دریافت کرده است. وی هم‌اکنون دانشیار گروه مهندسی برق دانشگاه خلیج فارس است. زمینه‌های پژوهشی مورد علاقه ایشان عبارت است از: سنجش از دور، پردازش تصاویر پزشکی، ماشین بینایی و هوش مصنوعی.

نشانی رایانامه ایشان عبارت است از:

a.keshavarz@pgu.ac.ir

[DOI: 10.61186/jsdp.20.3.197]