Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 9, Number 3, Fall 2020, pp. 214-228
Received: 07/25/2019, Revised: 03/27/2020, Accepted: 07/25/2020

A Hybrid Ensemble (Learning) of Knowledge-based Approaches for Content-based Filtering and Managing Information Resources

Morteza Jaderyan[1], Hassan Khotanlou[2*]
1- Department of Comuter Engineering, Bu-Ali Sina University, Hamadan, Iran.
2*- Department of Comuter Engineering, Bu-Ali Sina University, Hamadan, Iran.
[1]m.jaderyan92@basu.ac.ir, [2*]khotanlou@basu.ac.ir

Corresponding author address: Hassan Khotanlou, Department of Comuter Engineering, Faculty of Engineering, Bu-Ali Sina University, Hamadan, Iran.

Abstract- Knowledge-oriented content-based filtering techniques are among the most effective ways to search, filter, and manage information resources. In this paper, a novel filtering framework for (textual) information resource management purposes is introduced. The proposed method uses the collective knowledge of ontology and structured knowledge bases for developing semantic similarity methods. The semantic similarity methods are used to filter and classify documents in accordance with user preferences. Also, the knowledge-based semantic similarity methods are integrated in a “Mixture of Experts” model so that information resources and documents are filtered and managed based on the collective knowledge of these methods (experts). The integration of knowledge-based methods in the learning "Mixture of Experts" model is a novel idea and one of the main contribution of this paper. The evaluation results suggest that the integration of knowledge-based semantic similarity measures in "Mixture of Experts" model improves system performance and leads to the accurate classification of documents.

Keywords- Information filtering and management; content classifier; Semantic Similarity; Ontology; Ensemble Learning;

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
جلد ۹، شماره ۳، پائیز ۱۳۹۹ ، صفحه ۲۱۴-۲۲۸
دریافت: ۱۳۹۸/۰۵/۰۳، بازنگری: ۱۳۹۹/۰۱/۰۸، پذیرش: ۱۳۹۹/۰۵/۰۴

چارچوب ترکیبی (یادگیری-دانش‌محور) برای فیلتر محتوایی اطلاعات و مدیریت منابع اطلاعاتی

مرتضی جادریان[۱]، حسن ختن‌لو*[۲]

۱-دانشگاه بوعلی سینا، دانشکده مهندسی، گروه مهندسی کامپیوتر، همدان، ایران
*۲-دانشگاه بوعلی سینا، دانشکده مهندسی، گروه مهندسی کامپیوتر، همدان، ایران
[1]m.jaderyan92@basu.ac.ir, [2]*Khotanlou@basu.ac.ir

* نشانی نویسنده مسئول: همدان- خیابان مهدیه- دانشگاه بوعلی سینا- گروه کامپیوتر

چکیده- روش‌های فیلتر محتوایی مبتنی بر دانش، روش‌های مؤثری برای جستجو، فیلتر کردن و مدیریت اطلاعات هستند. در این مقاله، یک چارچوب بدیع فیلتر و مدیریت منابع اطلاعاتی متنی معرفی می‌شود. روش پیشنهادی از دانش جمعی/گروهی مدل شده در آنتولوژی و پایگاه‌های دانش ساخت‌یافته جهت توسعه روش‌های محاسبه معنایی شباهت استفاده می‌کند. از روش‌های محاسبه معنایی شباهت توسعه داده شده برای فیلتر و دسته‌بندی کردن اسنادی استفاده می‌شود که حاوی اطلاعات متنی منطبق با ترجیحات کاربری هستند. همچنین، روش‌های توسعه داده شده در یک مدل «ترکیب خبرگان» با یکدیگر یکپارچه می‌شوند تا تصمیمات مرتبط با فیلتر و مدیریت منابع اطلاعاتی، از طریق اجتماع دانش خبره‌ها اتخاذ گردد. یکپارچه‌سازی روش‌های مبتنی بر دانش در مدل یادگیری ماشین «ترکیب خبرگان» ایده بدیع پیشنهادی در این مقاله است. نتایج ارزیابی نشان می‌دهد اجماع خبرگی روش‌های مبتنی بر دانش در مدل یادگیری گروهی «ترکیب خبرگان» عملکرد سیستم را ارتقاء می‌بخشد و منجر به دسته‌بندی دقیق اسناد متنی می‌شود.

واژه‌های کلیدی: فیلتر و مدیریت اطلاعات؛ دسته‌بند محتوا؛ شباهت معنایی؛ آنتولوژی؛ یادگیری ترکیبی؛

۱- مقدمه

در یک دهه اخیر جستجو برای یافتن اطلاعات مرتبط با نیازهای اطلاعاتی کاربران به یک فعالیت روزمره تبدیل شده است. با این حال، یافتن منابع اطلاعاتی منطبق با ترجیحات کاربران فرآیندی چالش‌برانگیز است. فیلتر اطلاعات فرآیندی است که در نتیجه آن اطلاعات منطبق با ترجیحات کاربری شناسایی و به کاربر نمایش داده می‌شوند. سیستم‌های فیلتر محتوایی اطلاعات برای شناسایی محتوای اطلاعاتی مشابه با ترجیحات کاربری، معمولاً یک نمایش ایجاد شده از ترجیحات کاربری (نمایه کاربری) را با یک نمایش ایجاد شده دیگر از اسناد متنی (نمایه اسناد) مقایسه می‌کنند. این دسته از سیستم‌ها، تحلیل جامعی از محتوای اطلاعاتی اسناد متنی انجام می‌دهند و اسناد متنی را بر اساس میزان شباهت آن‌ها به ترجیحات کاربری فیلتر می‌کنند. روش‌های سنتی فیلتر محتوایی اطلاعات تنها از تحلیل ویژگی‌های رشته‌ای موجود در اسناد برای فیلتر اسناد متنی استفاده می‌کنند. با این حال، سیستم‌های نوین از تکنیک‌های هوشمند یادگیری ماشین و مبتنی بر دانش برای استخراج ویژگی‌ها از منابع اطلاعاتی و دسته‌بندی آن‌ها بهره می‌برند [۱، ۲]. شناخت نیازهای

دقیق اطلاعاتی کاربران و تطابق آن‌ها با محتوای اسناد متنی، وظیفه چالش‌برانگیز برای سیستم‌های پردازش و مدیریت اطلاعات است. دیگر مشکل شایع چنین سیستم‌هایی، طبقه‌بندی‌های از پیش تعیین شده برای ترجیحات کاربری است؛ به عبارت دیگر، کاربرانی که اشتراکات معنی‌داری با ترجیحات از پیش تعیین شده نداشته باشند، اسناد متنی مطابق با نیازهای اطلاعاتی آن‌ها توسط سیستم قابل شناسایی نخواهد بود. ظهور وب معنایی سبب ایجاد روش‌های مبتنی بر آنتولوژی جهت نمایش محتوا شد. در طی یک دهه اخیر، آنتولوژی‌ها به عنوان ابزاری مفید برای یکپارچه‌سازی دانش در سیستم‌های فیلتر و مدیریت اطلاعات شناخته شده‌اند [۳]. از آنتولوژی می‌توان برای غلبه بر مشکل ابهام در محتوای متنی و بهبود نمایش آن‌ها استفاده کرد [۴]. پایگاه‌های دانش نظیر ویکی‌پدیا [۶،۵،۴] منبع غنی از اطلاعات دامنه محسوب می‌شوند و یکپارچه‌سازی آن‌ها در سیستم‌های فیلتر اطلاعات، نمایش ایجاد شده از نمایه‌های کاربری و نمایه‌های اسناد را بهبود می‌بخشد و تطابق میان نمایه‌ها را تسهیل می‌کند. همچنین، یکپارچه‌سازی دانش ساخت‌یافته‌ی پایگاه‌های دانش و آنتولوژی سبب مدل‌سازی بهتر محتوای متنی و ترجیحات کاربری می‌شود.

ایده ارائه شده در این مقاله را می‌توان به این شکل خلاصه کرد: در مرحله اول، نمایه‌های کاربری و اسناد تشکیل می‌شوند. نمایه‌های کاربری از طریق نمایش ترجیحات صریح کاربری و یا اسنادی که بهترین شکل ترجیحات کاربری را منعکس می‌کنند، تشکیل می‌شوند. نمایه اسناد از طریق توکن کردن اسناد و وزن‌دهی به مفاهیم ساخته می‌شود. سپس، نمایه‌ها پیش‌پردازش می‌شوند تا ویژگی‌های حاوی اطلاعات مفید شناسایی شوند. ویژگی‌های شناسایی شده ابهام‌زدایی معنایی می‌شوند. سپس، نمایه‌ها جهت بهبود نمایش تولید شده غنی‌سازی می‌شوند. در مرحله بعد، شباهت میان نمایه‌ها از طریق روش‌های دانش‌محور شباهت معنایی محاسبه می‌شود و تصمیم برای فیلتر کردن یک سند دلخواه داده شده، توسط هر کدام از خبرگان شباهت معنایی اتخاذ می‌گردد. در مرحله آخر، تصمیم گروهی/جمعی در رابطه با فیلتر سند با استفاده از مدل یادگیری گروهی اتخاذ می‌شود. بر اساس تصمیم گروهی/جمعی خبرگان، اسناد ورودی به سیستم در دسته «مشابه/مرتبط» یا «نامرتبط/نامشابه» طبقه‌بندی می‌شوند.

این مقاله به این شکل سازمان‌دهی شده است: در بخش دوم، مطالعات و تحقیقات مرتبط بررسی می‌شوند. در بخش سوم، ساختار آنتولوژی و پایگاه‌های دانش استفاده شده بررسی می‌شوند. در بخش چهارم، روش ترکیبی (یادگیری ماشین-دانش‌محور) پیشنهادی برای فیلتر محتوایی اطلاعات شرح داده خواهد شد. در بخش پنجم، نتایج ارزیابی روش پیشنهادی نمایش داده می‌شوند و در بخش ششم، جمع‌بندی و نتیجه‌گیری ارائه خواهند شد.

۲- مطالعات مرتبط

عمده نواقص موجود در سیستم‌های فیلتر و مدیریت اطلاعات از ابهام موجود در محتوا و ناتوانی سیستم در شناخت معنای موجود در آن‌ها نشاءت می‌گیرند. یکی از مهم‌ترین زمینه‌های تحقیقاتی برای بررسی و مرتفع کردن این نقص، توسعه سیستم‌های هوشمند مبتنی بر یادگیری مدیریت اطلاعات است. با این حال، روش‌های مبتنی بر یادگیری به داده‌های یادگیری وابسته هستند و کوچک‌ترین تغییرات در دامنه اطلاعاتی باعث کاهش عملکرد و دقت آن‌ها می‌شود. در سال‌های اخیر، روش‌های مبتنی بر دانش توجه بسیاری از محققان را به خود جلب کرده است. این روش‌ها از منابع خارجی دانش (آنتولوژی و پایگاه دانش) برای نگاشت دانش دامنه یا سطح بالا به سیستم‌های اطلاعاتی استفاده می‌کنند. بر خلاف روش‌های مبتنی بر یادگیری ماشین، روش‌های دانش‌محور نیازی به یادگیری ندارند و تغییرات در دامنه اطلاعاتی، تأثیر شگرف بر عملکرد کلی سیستم ندارد. با این حال، این روش‌ها به زمان محاسباتی بیشتری نسبت به روش‌های مبتنی بر یادگیری ماشین نیاز دارند.

مدل‌های یادگیری برای نمایش محتوا و دسته‌بندی اسناد نیز استفاده می‌شوند. نویسندگان در [۷] روش Bag-of-Concepts را به عنوان یک روش نمایش محتوای اسناد پیشنهاد می‌کنند که خوشه‌های مفاهیم موجود در اسناد را از طریق خوشه‌بندی بردارهای کلمه تولید شده از Word2Vec تولید و از تناوب خوشه‌های مفاهیم برای نمایش بردارهای اسناد متنی استفاده می‌کند. در مرجع [۸]، از یک مدل معنایی نهان جدید برای یادگیری نمایش برداری‌های معنایی معرفی شده که ویژگی‌های N-gram توسط این مدل شناسایی می‌شوند و سپس، به منظور ایجاد بردار ویژگی در سطح جمله با یکدیگر تجمیع می‌شوند. در [۹]، از یک مدل شبکه عصبی برای دسته‌بندی اسناد استفاده شده که هر کلمه موجود در یک جمله را دریافت می‌کند، اطلاعات نهان موجود در آن را استخراج و به یک بردار معنایی تبدیل می‌کند. از بردارهای تولید شده برای دسته‌بندی اسناد استفاده می‌شود.

دانش ساخت‌یافته موجود در آنتولوژی و پایگاه‌های دانش، ابزار مناسبی جهت استخراج معنای نهان در محتویات فراهم می‌کند که آن‌ها را

برای استفاده در فیلتر اطلاعات مناسب می‌کند. میبدل و همکارانش یک روش فیلتر محتوایی برای استفاده در پروژه روزنامه الکترونیک شخصی پیشنهاد داده‌اند [۱، ۱۰]. در این رویکرد، ترجیحات کاربری و محتویات اسناد توسط مفاهیم استنتاج شده از آنتولوژی سه-سطحی به نام IPTC_NewsCode نمایش داده می‌شود. در [۱۱] یک سیستم استخراج و بازیابی اطلاعات مبتنی بر آنتولوژی در حوزه فوتبال ارائه شده است. در [۱۲] یک روش مبتنی بر آنتولوژی برای بازیابی اسناد متنی پیشنهاد شده است. در این روش، از آنتولوژی دامنه برای ایجاد روابط معنایی بین مفاهیم موجود در اسناد استفاده می‌شود. سپس، یک مکانیسم وزن‌دهی و یک روش محاسبه شباهت که مبتنی بر روابط معنایی است، شباهت یک سند را به پرس‌وجوی کاربر مشخص می‌کند. در [۱۳] یک روش شخصی‌سازی شده جهت جستجو و بازیابی اسناد متنی معرفی شده است. در این مقاله، اسناد متنی توسط نگاشت مفاهیم به یک ساختار گراف مانند نمایش داده می‌شوند. روابط میان مفاهیم نیز با استفاده از یک آنتولوژی تحت وب به نام ODP شناسایی می‌شوند.

تحقیقات زیادی در مورد استفاده از ویکی‌پدیا برای بهبود نمایش اسناد یا پرس‌وجوی کاربر و همچنین ارتقاء عملکرد مدل‌های بازیابی اطلاعات انجام شده است [۱۴، ۱۵]. گابریوولیچ و همکارانش یک روش محاسبه شباهت متون با استفاده از تجزیه و تحلیل معنای صریح مبتنی بر ویکی‌پدیا معرفی کرده‌اند [۱۶]. آن‌ها یک مترجم معنایی برای شاخص‌گذاری مفاهیم معرفی کردند. از مترجم معنایی به عنوان یک دسته‌بند استفاده می‌شود و شاخص‌های تولید شده رتبه‌بندی خواهند شد. مالو و همکاران استفاده از ویکی‌پدیا (به عنوان یک منبع دانش پوشش دهنده چندین دامنه اطلاعاتی مختلف) را برای محاسبه معنایی شباهت میان مفاهیم مورد مطالعه قرار دادند [۱۵]. در [۱۷] یک روش جدید محاسبه معنایی شباهت پیشنهاد شده است. این روش از ویژگی‌های اطلاعاتی موجود در ویکی‌پدیا مانند مقالات، گراف دسته‌بندی مقالات و اطلاعات معنایی ویکی‌پدیا استفاده می‌کند. روش پیشنهادی در [۱۸] هر سند را به عنوان یک بردار مفهومی در فضای معنایی ویکی‌پدیا نمایش می‌دهد تا معانی موجود در اسناد متنی را مدل کند.

گائو و همکاران روش جدیدی برای محاسبه شباهت معنایی بر اساس گراف معنایی و نظریه محتوای اطلاعاتی ارائه می‌دهند [۱۹]. روش پیشنهادی از طریق یک تبدیل غیرخطی، طول وزن‌دار کوتاه‌ترین مسیر بین مفاهیم را به امتیاز مشابهت میان آن‌ها تبدیل می‌کند.

اگرچه پایگاه دانش BNC به‌طور گسترده‌ای در کاربردهای پردازش زبان طبیعی مورد استفاده قرار می‌گیرد اما هیچ‌گونه شواهدی در مورد استفاده از این پایگاه دانش برای پیاده‌سازی یک سیستم فیلتر محتوایی در تحقیقات قبلی یافت نشده است. با توجه به مطالعات ما، این مقاله اولین مورد استفاده از پایگاه دانش BNC برای مقاصد فیلتر محتوایی اسناد متنی است.

۳- ساختار آنتولوژی و پایگاه‌های دانش

پیش از معرفی سیستم پیشنهادی برای فیلتر و مدیریت اطلاعات، ساختار آنتولوژی و پایگاه‌های دانش استفاده شده در این تحقیق به‌طور خلاصه شرح داده می‌شوند.

آنتولوژی: آنتولوژی OntoWordNet (حاصل هم‌ترازی OWL از آنتولوژی WordNet و کتابخانه آنتولوژی DOLCE-Lite plus) در روش پیشنهادی استفاده شده است. هر مفهوم موجود در آنتولوژی به شکل مجموعه‌ای از مفاهیم هم‌معنی سازمان‌دهی شده است تا مفاهیم مشابه (از نظر محتوای اطلاعاتی) قابل شناسایی باشند. چنین ساختاری فرآیند غنی‌سازی محتوا را تسهیل می‌کند [۲۰]. این آنتولوژی ماحصل پروژه تحقیقاتی موسسه علوم و فنّاوری شناختی وابسته به شورای پژوهشی ملی ایتالیا است [۲۰].

Wikipedia and BNC: داده‌های ویکی‌پدیا و BNC که در این مقاله مورد استفاده قرار گرفته شده‌اند، جهت استفاده دانشگاهی، از طریق پروژه D.I.S.C.O در دسترس هستند. داده‌های نامبرده شده دارای ساختار یکسان هستند. روش تولید این مجموعه داده در مراجع [۲۱، ۲۲] شرح داده شده‌اند. هر دو ساختار داده شامل دو مجموعه داده است: (۱) بردار کلمه مرتبه اول که شامل کلماتی است که در کنار هم در ساختار ویکی‌پدیا و BNCظاهر می‌شوند و (۲) بردار کلمه مرتبه دوم که شامل کلماتی است که در قالب‌های محتوایی مشابه رخ می‌دهد و از نظر معنایی قابل جایگزینی با یکدیگر هستند.

WordNet: یک آنتولوژی معنایی برای زبان انگلیسی است. ساختار اصلی WordNet را مجموعه‌های مفاهیم هم‌معنی تشکیل می‌دهد. جزئیات بیشتر در مورد WordNet در مرجع [۲۳] در دسترس است.

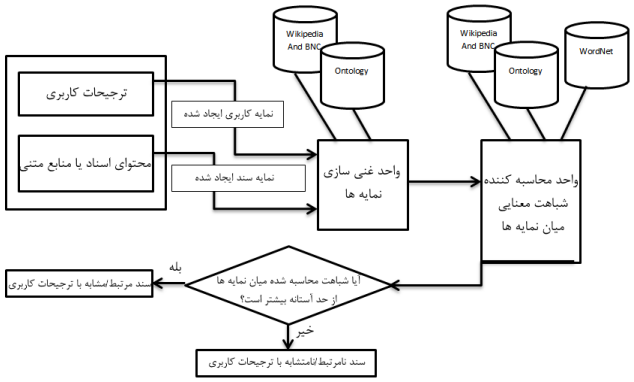
۴- روش پیشنهادی

نمایی از چارچوب کلی سیستم ترکیبی متشکل از روش‌های یادگیری ماشین-دانش محور در شکل ۱ نشان داده شده است. ورودی سیستم شامل ترجیحات کاربری و اسناد/ صفحات وب است و خروجی آن

مقدار شباهتی است که نشانگر شباهت نمایه کاربری به نمایه اسناد است.

۴-۱- نمایه کاربری و نمایه سند

به منظور ارائه تجربه شخصی‌سازی شده، نمایه‌های جداگانه‌ای به کاربران اختصاص داده می‌شود. هر نمایه، ترجیحات کاربری یک فرد خاص را نشان می‌دهد. نمایه‌های کاربری در مراحل اولیه ایجاد و به صورت آفلاین به‌طور مرتب به‌روزرسانی می‌شوند. در ابتدا، نمایه کاربری توسط فهرستی از مفاهیم که ترجیحات کاربر را نمایش می‌دهند، نمایش داده می‌شوند. یک مقدار در بازه [۰،۱] (وزن) به هر مفهوم اختصاص داده می‌شود که میزان علاقه کاربر به هر مفهوم را نشان می‌دهد. درصورتی‌که ترجیحات کاربری از اسناد متنی مورد علاقه و مشابه ترجیحات کاربری استخراج شوند، مکانیسم وزن‌دهی TF-IDF اهمیت هر مفهوم را تعیین می‌کند.



شکل ۱ – چارچوب کلی سیستم ترکیبی پیشنهادی

گام بعدی ساختن نمایه‌های اسناد است. اعمال پیش‌پردازشی زیر پس از بارگذاری اسناد در سیستم انجام می‌شوند:

- (۱) حذف کلمات Stop-word، (۲) پردازش bi-gram و Uni-gramکلمات، (۳) برچسب زدن Part of Speech (POS) کلمات، (۴) lemmatization و (۵) شناسایی named-entityها.

برای برطرف کردن ابهام موجود در مفاهیم، یک روش «تمیز معانی کلمات» که از تحقیق ارائه شده در [۲۴] الهام گرفته شده، معرفی شده است. فرضیه اصلی این روش این است که معانی مشابه در قالب‌های محتوایی مشابه رخ می‌دهند. برای این کار، مراحل زیر انجام می‌شود:

۱) پنجره محتوایی به اندازه $7^{\pm}$ در اطراف مفهوم مورد نظر ایجاد می‌شود. همچنین، بردار کلمه مرتبه اول متناظر با هر عضو پنجره متنی ایجاد شده بازیابی می‌شود. تجمیع پنجره محتوا و بردارهای کلمه بازیابی شده، یک «بردار محتوا» به ازای هر مفهوم می‌سازد.

۲) تمامی معانی ممکن مفهوم مورد نظر، نمونه به‌کارگیری معانی در یک جمله و تعریف خلاصه شده از معانی (Gloss) با استفاده از پایگاه دانش WordNet استخراج می‌شوند. چنین فرایندی سبب ایجاد «بردار معنی» برای هر کدام از معانی ممکن یک مفهوم خواهد شد. همچنین بردار کلمه مرتبه اول متناظر با هر عضو بردارهای معنی تشکیل شده بازیابی شده و به بردار معنی مورد نظر اضافه می‌شوند. محاسبه میزان شباهت میان بردار محتوا و هر کدام از بردارهای معنی مشخص می‌کند که معنی صحیح هر مفهوم در قالب محتوایی دربرگیرنده آن کدام است.

۳) ترکیب وزن‌دار خطی از معیارهای شباهت Jaro-Winkler و کسینوسی برای محاسبه میزان شباهت استفاده می‌کند.

$$Sim(Sense_{Vctor}, context_{Vector}) = \frac{1}{2}(Cosine_{Sim}(Sense_{Vctor}, context_{Vector}) + Jaro\_winkler_{Sim}(Sense_{Vctor}, context_{Vector})) \quad (1)$$

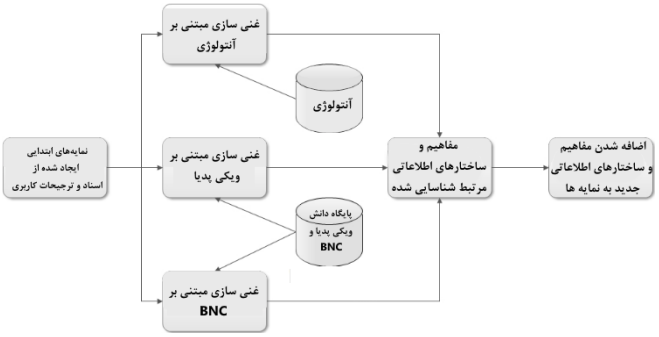
۴) بردار معنی که بیشترین شباهت را به بردار محتوا داشته باشد به‌عنوان بردار معنی صحیح شناخته می‌شود و معنای متناظر با این بردار جهت حاشیه‌نویسی مفاهیم استفاده می‌شود. سپس، مفاهیم با استفاده از روش TF-IDF وزن‌دهی می‌شوند. خروجی این مرحله یک نمایه سند متشکل از مجموعه‌ای از مفاهیم وزن‌دار است. تمامی مفاهیم به وسیله معنی واقعی آن‌ها حاشیه‌نویسی شده‌اند.

۴-۲- غنی‌سازی نمایه‌ها

محتوای غنی شده، نمایش ایجاد شده از نمایه کاربر و سند را بهبود می‌بخشد. واحد غنی‌سازی محتوا از دانش ساخت‌یافته آنتولوژی و پایگاه‌های Wordnet، ویکی‌پدیا و BNC بهره می‌برد تا ترجیحات کاربری جدیدی که ممکن است توسط کاربر از قلم افتاده باشد را کشف نماید. همچنین، این واحد نقش مهمی در پیدا کردن مفاهیمی دارد که محتوای اطلاعاتی یک سند را بهبود می‌بخشند. فرآیند غنی‌سازی محتوا در شکل ۲ نشان داده شده است.

مفاهیم غنی شده استنتاجی برای نمایه کاربر به‌طور مستقیم به این نمایه اضافه می‌شوند. با این حال، از آنجایی که تعداد مفاهیم غنی شده استنتاجی برای نمایه سند بالا است، این مفاهیم در حافظه موقتی به

نام «مفاهیم مرتبط» ذخیره می‌شوند تا در مرحله نهایی ایجاد نمایه سند، پردازش تکمیلی روی آن‌ها انجام شود.



شکل ۲ – چارچوب کلی سیستم ترکیبی پیشنهادی فرآیند غنی‌سازی محتوا

غنی‌سازی مبتنی بر آنتولوژی: در این مرحله، از دو تکنیک مبتنی بر آنتولوژی به نام‌های set spreading [۲۵] و domain-based inference: upward Programming [۲۶] برای غنی‌سازی محتوا و بهبود نمایش نمایه‌ها استفاده می‌شود. در روش Set Spreading، از ویژگی سامان‌دهی مفاهیم متشابه از نظر قالب محتوایی در ساختار OntoWordNet استفاده می‌شود تا مفاهیم مشابه به هر مفهوم در نمایه‌های کاربری و سند شناسایی شود. روش Upward Programming بر این ایده استوار است که اگر یک کاربر به یک مفهوم در ساختار آنتولوژی علاقه‌مند باشد، نسبت به والد/فرزندان مستقیم مفهوم در ساختار سلسله مراتبی نیز به‌طور نسبی علاقه‌مند است. غنی‌سازی مبتنی بر آنتولوژی در مدل‌سازی ترجیحات کاربر بسیار مؤثر شناخته شده است [۲۷]. ما از این ایده برای غنی‌سازی محتوای نمایه‌ها توسط مفاهیم والد/فرزند در ساختار سلسله مراتبی آنتولوژی استفاده کردیم. مفاهیم غنی‌شده توسط رابطه زیر وزن‌دهی می‌شوند و به نمایه‌های مربوطه اضافه می‌شوند:

$$related\_concept\_weight = original\_concept\_weight * 0.8; \quad (2)$$

ضریب ۰.۸ از آزمایش‌های انجام شده روی داده‌ها به دست آمده است.

غنی‌سازی مبتنی بر ویکی‌پدیا: در این بخش، از بردارهای مرتبه اول و مرتبه دوم ویکی‌پدیا برای یافتن مفاهیم هم اتفاق و مفاهیم مشابه با هر یک از مفاهیم ظاهر شده در نمایه‌های کاربری و اسناد استفاده می‌شود. مفاهیم غنی شده توسط روابط زیر وزن‌دهی می‌شوند و به نمایه‌های مربوطه اضافه می‌شوند:

$$related\_second\_order\_concept\_weight = original\_concept\_weight * 0.7; \quad (3)$$

$$related\_first\_order\_concept\_weight = original\_concept\_weight * 0.5; \quad (4)$$

ضرایب ۰.۷و ۰.۵ با آزمایش‌های مشخصی روی داده‌ها به دست آمده‌اند.

غنی‌سازی مبتنی بر BNC: در این تحقیق، تنها از بردار کلمه مرتبه دوم BNC برای پیدا کردن مفاهیم مشابه با هر یک از مفاهیم ظاهر شده در نمایه‌های کاربری و اسناد استفاده می‌شود. . مفاهیم غنی شده توسط رابطه زیر وزن‌دهی می‌شوند و به نمایه‌های مربوطه اضافه می‌شوند:

$$related\_most\_similar\_concept\_weight= original\_concept\_weight * 0.6; \quad (5)$$

ضریب ۰.۶ توسط آزمایش‌های انجام شده روی داده‌ها به دست آمده است.

۴-۳- مرحله نهایی تولید نمایه سند

خروجی فرآیند غنی‌سازی یک ساختار داده‌ای به نام «مفاهیم مرتبط» است. این ساختار داده‌ای شامل مفاهیم غنی شده مرتبط با نمایه سند است. با این حال، اندازه این ساختار داده‌ای با افزایش تعداد مفاهیم در نمایه سند آن به‌طور نمایی رشد می‌کند. بنابراین، فرآیند فیلتر کردن اطلاعات از نظر محاسباتی مقرون به صرفه نخواهد بود. به منظور کاهش اندازه این ساختار داده، از روش محاسبه معنایی شباهت Resnik و یک آستانه مشابهت (α) استفاده می‌شود. به عبارت دیگر، تنها مفاهیمی که یکی از شرایط زیر را برآورده می‌کنند به نمایه سند اضافه می‌شوند:

۱) مفهوم غنی شده‌ای که مطابقت کامل (رشته‌ای) با یکی از مفاهیم در نمایه کاربری دارد.
۲) مفهوم غنی شده‌ای که امتیاز مشابهت آن با یک مفهوم در نمایه کاربری از α بیشتر است (مقدار شباهت با استفاده از معیار مشابهت Resnik محاسبه می‌شود).
۳) به ازای هر مفهوم در نمایه سند، سه مفهوم غنی‌شده‌ای که بیشترین امتیاز مشابهت را با مفهوم مربوطه در نمایه سند دارند، به این نمایه اضافه می‌شوند.

پارامتر α یک پارامتر کنترلی است. این پارامتر اندازه نمایه سند را کنترل می‌کند. همان‌طور که مقدار α افزایش می‌یابد، اندازه نمایه سند

کاهش می‌یابد و برعکس. بهترین مقدار برای α مقدار ۰.۶ است که از طریق آزمون و خطا به دست آمده است.

۴-۴- روش‌های فیلتر اطلاعات مبتنی بر شباهت معنایی

یکی از مؤلفه‌های اصلی سیستم ترکیبی فیلتر و بازیابی اطلاعات پیشنهاد شده، واحد محاسبه معنایی شباهت (مبتنی بر دانش) است. وظیفه این واحدها، تشخیص سطح تشابه بین ترجیحات کاربری و محتویات اسناد است. در اینجا چهار نوع رویکرد مبتنی بر دانش را برای محاسبه سطح معنایی مشترک میان نمایه‌ها مطالعه کرده‌ایم. دلیل چنین کاری مطالعه اثر یکپارچه‌سازی آنتولوژی و پایگاه‌های دانش بر عملکرد و دقت روش پیشنهادی است. دلیل دیگر این است که مشخص کنیم آیا یکپارچه‌سازی روش‌های دانش‌محور با چارچوب یادگیری گروهی پیشنهادی می‌تواند باعث دسته‌بندی صحیح نمونه‌هایی شود که پیش از این توسط روش‌های دیگر به اشتباه دسته‌بندی می‌شدند؟

در این مقاله یک چارچوب یادگیری گروهی به نام «ترکیب خبرگان» معرفی شده است. این چارچوب، واحدهای محاسبه معنایی شباهت را با روش‌های مبتنی بر یادگیری گروهی یکپارچه می‌کند. در نهایت، اسناد بر اساس امتیاز شباهت میان نمایه‌ها و یک پارامتر کنترلی به نام β فیلتر می‌شوند. پارامتر β حساسیت سیستم را در تطابق اسناد با ترجیحات کاربری کنترل می‌کند. مقدار β بالا به معنای فیلتر دقیق‌تر و در عین حال، سخت‌گیرانه‌تر اطلاعات است.

فیلتر اطلاعات مبتنی برآنتولوژی: برای محاسبه معنایی شباهت میان نمایه‌ها، دو ویژگی مهم آنتولوژی OntoWordnet در نظر گرفته شده است: (۱) سازمان‌دهی مفاهیم مشابه به‌صورت مجموعه‌های هم‌معنی و (۲) روابط معنایی و سلسله مراتبی ایجاد شده میان مفاهیم؛ به‌طور خاص روابط والد/فرزند و رابطه هم‌معنی. در مرحله اول مفاهیم موجود در هر دو پروفایل را به آنتولوژی OntoWordNet نگاشت می‌شود. برای کاهش بار محاسباتی سیستم و نادیده گرفتن مفاهیم بسیار عمومی یا بسیار خاص در فرآیند محاسبه معنایی شباهت، تنها مفاهیم موجود در نمایه‌ها، مفاهیم والد/ فرزند و مفاهیم اجدادی/ نوادگانی در نظر گرفته شده‌اند. بنابراین، با سه نوع مشابهت روبرو خواهیم شد:

- تشابه کامل: اگر یک مفهوم در هر دو نمایه ظاهر شود، یا اگر یک مفهوم در یکی از نمایه‌ها ظاهر شود و معادل آن (مفهوم مشابه از نظر قالب محتوایی) در نمایه دیگر ظاهر شود.
- تشابه نزدیک: یک مفهوم در یک نمایه ظاهر می‌شود، در حالی که فرزند/ والد این مفهوم یا مفهوم هم‌معنی با فرزند/ والد آن در نمایه دیگر ظاهر شود.
- تشابه جزئی: یک مفهوم در یک نمایه ظاهر شود، در حالی که نواده/ جد آن یا مفهوم هم‌معنی با نواده/ جد آن در نمایه دیگر ظاهر شود.

این مکانیسم (شکل ۳) به چارچوب فیلتر اطلاعات اجازه می‌دهد تا ظاهر شدن مفاهیم اصلی و همچنین مفاهیم مرتبط با محتوا در نمایه‌ها را در هنگام محاسبه معنایی شباهت لحاظ کند.

با Ontological Similarity Measure (OSM) میزان شباهت دو نمایه بر اساس تعداد مشابهت‌های کامل، نزدیک و جزئی میان مفاهیم ظاهر شده در هر دو نمایه و همچنین، وزن اختصاص داده شده به مفاهیم موجود در پروفایل کاربری سنجیده می‌شود:

$$OSM = \frac{\sum_{i,j \in U,D} w_i * Score_{j,i}}{\sum_{i,j \in U,D} w_i} \qquad (6)$$

که "i" و "j" به ترتیب اندیس مفاهیم در نمایه کاربر و نمایه سند، "D" مجموعه مفاهیم نمایه سند، "U" مجموعه مفاهیم در نمایه کاربر و "$Score_{j,i}$" مقدار شباهت نهایی برای هر جفت مفهوم است.

```
w= weight of concept in user profile
Score=0
Begin
Begin
For each concept (D) in document profile:
   - If D (or equivalent concept to D)is in user profile then
     score+=(1*w);
   - If parent (or equivalent concept to parent)of D is in user
     profile, then score+=((2/3)*w);
   - If grandparent(or equivalent concept to grandparent) of D
     is in user profile, then score+=((2/5)*w);
End
Begin
For each concept (U) in User profile:
   - If parent(or equivalent concept to parent) of U is in
     document profile then score+=((2/5)*w));
   - If grandparent(or equivalent concept to grandparent) of U
     is in document profile then score+=((1/5)*w);
End
End.
```

شکل ۳ - الگوریتم فیلتر اطلاعات مبتنی بر آنتولوژی

فیلتر اطلاعات مبتنی بر ویکی‌پدیا و BNC: با توجه به یکسان بودن ساختار اطلاعاتی ویکی‌پدیا و BNC، روش یکسانی برای محاسبه

شباهت معنایی استفاده شده است. مزیت اصلی این استراتژی این است که می‌توان نتایج ارزیابی دو روش را مقایسه کرد و مشخص کرد کدام پایگاه دانش برای مقاصد فیلتر و بازیابی اطلاعات مناسب‌تر است. برای این منظور، معیار شباهت مبتنی بر بردار مرتبه اول و معیار شباهت مبتنی بر بردار مرتبه دوم توسعه داده شده است.

برای محاسبه شباهت معنایی بین دو مفهوم داده شده، بردار مرتبه اول مرتبط با هر مفهوم با استفاده از کتابخانه Apache Lucene بازیابی و شباهت میان دو مفهوم، از طریق مقایسه این بردارها محاسبه می‌شود. برای محاسبه شباهت میان دو مفهوم، از روش الهام گرفته شده از نظریه اطلاعاتی ارائه شده توسط Lin [۲۸، ۲۱، ۲۲] استفاده می‌شود:

$$score = \frac{\sum_{(rel,w')} freq(A,*_{rel},*_w) + freq(B,*_{rel},*_w)}{\sum_{(rel,w')} freq(A,*_{rel},B) + \sum_{(rel,w')} freq(B,*_{rel},A)} \quad (7)$$

$* \; rel = \{contextually\_similar\ rlation\}$

$* \; w = \{concepts\ in\ either\ the\ document\ or\ user\ profile\}$

در این رابطه، Aو B مفاهیم ظاهر شده در نمایه‌های اسناد و کاربری هستند. تابع *freq()* تناوب مفاهیم A وB در رابطه هم‌اتفاقی تعریف شده میان مفاهیم را محاسبه و بیان می‌کند که شباهت میان دو مفهوم A و B بر اساس مشترکات و تفاوت‌های میان ساختارهای اطلاعاتی آن‌ها (بردارهای مرتبه اول دو مفهوم) محاسبه می‌شود.

برای محاسبه شباهت معنایی بین دو مفهوم، بردار مرتبه دوم مربوط به هر مفهوم با استفاده از کتابخانه Apache Lucene بازیابی و شباهت میان دو مفهوم از طریق مقایسه محتوای اطلاعاتی این بردارها محاسبه می‌شود. برای محاسبه شباهت میان دو مفهوم از روش الهام گرفته شده از نظریه اطلاعاتی ارائه شده توسط Lin [۲۸، ۲۱، ۲۲] استفاده می‌شود:

$$score = \frac{\sum_{(rel,w')} freq(A,*_{rel},*_w) + freq(B,*_{rel},*_w)}{\sum_{(rel,w')} freq(A,*_{rel},B) + \sum_{(rel,w')} freq(B,*_{rel},A)} \quad (8)$$

$* \; rel = \{contextually\_similar\ rlation\}$

$* \; w = \{concepts\ in\ either\ the\ document\ or\ user\ profile\}$

در این رابطه، A و B مفاهیم ظاهر شده در نمایه‌های اسناد و کاربری هستند و تابع *freq()* تناوب مفاهیم A و B در روابط هم‌معنی تعریف شده میان مفاهیم را محاسبه می‌کند. این رابطه مشترکات و تفاوت‌های میان بردارهای مرتبه دوم دو مفهوم A و B را محاسبه می‌کند.

در اینجا، از مفهوم شباهت معنایی میان دو مفهوم برای محاسبه شباهت کلی میان نمایه کاربری و سند استفاده می‌شود که الگوریتم آن در شکل ۴ نشان داده شده است.

w_j: weight of concept in user profile.
w_i: weight of concept in document profile.
Score: semantic similarity between two concepts
Simval: aggregated score.
Value: Final value of similarity between profiles using Wikipedia or BNC.
Flag=false; Value=0; simval=0;
Begin
Begin
For each concept in user profile (j)
For each concept in document profile (i)
- *If perfect match happens then score=1;*
- *Else Compute semantic similarity between two concepts i and j (score).*
- *simval+=(score*(w_i*w_j));*

End
Value= $\frac{simval}{\sum_{i*j} w_i * w_j}$;
End.

شکل ۴ – الگوریتم فیلتر اطلاعات مبتنی بر ویکی‌پدیا و BNC

فیلتر اطلاعات مبتنی بر WordNet: پنج معیار محاسبه معنایی شباهت بر اساس دانش ساخت‌یافته WordNet پیاده‌سازی شده است. دو معیار Lin و Jiang & Conrath مبتنی بر مفهوم «محتوای اطلاعاتی کوچک‌ترین زیرمجموعه مشترک» [۲۳] پیاده‌سازی شده‌اند. محتوای اطلاعاتی، معیاری برای اندازه‌گیری خاصیت اطلاعاتی نهفته در یک مفهوم است و کوچک‌ترین زیرمجموعه مشترک دو مفهوم A و B، خاص‌ترین مفهومی است که در ساختار سلسله مراتبی آنتولوژی، جد این دو مفهوم محسوب می‌شود. سه روش دیگر، شباهت معنایی را بر اساس طول مسیر بین دو مفهوم در ساختار سلسله مراتبی WordNet محاسبه می‌کنند. این سه روش، روش‌های Path، Wu & Palmerو Leacock & Chodorow هستند. الگوریتم محاسبه معنایی شباهت مبتنی بر WordNet در شکل ۵ نشان داده شده است.

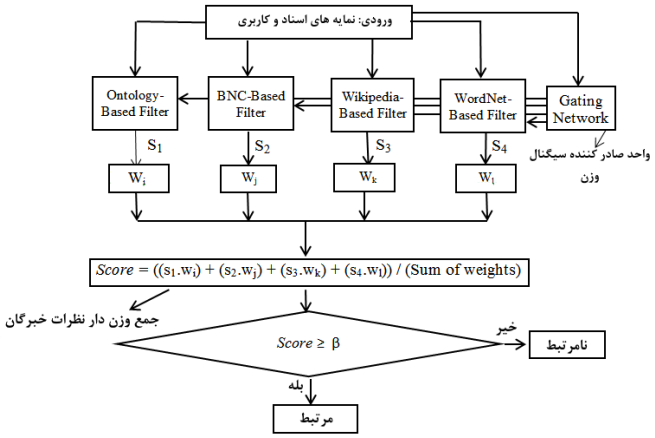
w_j: weight of concept in user profile.
w_i: weight of concept in document profile.
Score: semantic similarity between two concepts
Simval: aggregated score.
Value: Final value of similarity between profiles using WordNet.
Flag=false; Value=0; simval=0;
Begin
Begin
For each concept in user profile (j)
For each concept in document profile (i)
- *If perfect match happens then score=1;*
- *Else Compute semantic similarity between two concepts i and j (score).*
- *simval+=(score*(w_i*w_j));*

End
Value= $\frac{simval}{\sum_{i*j} w_i * w_j}$;
End.

شکل ۵ – الگوریتم فیلتر اطلاعات مبتنی بر WordNet

۴-۵- یادگیری ترکیبی (ترکیب خبرگان) برای فیلتر اطلاعات

نوآوری معرفی شده که روش ارائه شده را منحصر به فرد می‌کند، ایده یکپارچه‌سازی دانش ساخت‌یافته آنتولوژی و پایگاه‌های دانش با یک مدل یادگیری ترکیبی (گروهی) به نام «ترکیب خبرگان» (چارچوب شکل ۶) است. تخصص (خبره) جمعی روش‌های دانش‌محور محاسبه معنایی شباهت، تصمیم نهایی در مورد فیلتر کردن یک سند را مشخص می‌کند. تفاوت اصلی بین ساختار ترکیبی پیشنهادی و دیگرهای ساختارهای ترکیبی این است که روش‌های دانش‌محور در این ساختار ترکیب می‌شوند و نه روش‌های یادگیری ماشین. روش‌های دانش‌محور بر خلاف تکنیک‌های یادگیری ماشین به داده‌های آموزشی وابسته نیستند؛ اگر چه "Gating Network" هنوز به داده‌های دامنه برای آموزش متکی است تا بتواند وزن مناسب را به هر کدام از خبره‌ها اختصاص دهد.



شکل ۶ – چارچوب کلی مدل یادگیری ترکیبی «ترکیب خبرگان»

در این شکل، S_1، S_2، S_3 و S_4 مجموعه‌ای نمادین از مقادیر مشابهت‌های معنایی محاسبه شده به وسیله روش‌های دانش‌محور (خبرگان) و W_1، W_2، W_3 و W_4 نیز مجموعه‌ای نمادین از وزن‌های اختصاص داده شده به این روش‌ها (خبرگان) توسط "Gating Network" است. اگر به چارچوب یادگیری ترکیبی (گروهی) «ترکیب خبرگان» بیشتر دقت کنیم، ساختار این مدل را می‌توان به شکل زیر توصیف کرد:

- در این مدل، مجموعه‌ای از خبرگان $\{e_i(x)\}_{i=1}^{k}$ وجود دارند که هر خبره $e_i(x)$ یک روش دسته‌بند معنایی محتوا (روش دانش‌محور محاسبه معنایی شباهت) است.
- هر خبره یک بردار خروجی تولید می‌کند. این بردار احتمال اینکه یک سند به کلاس 'relevant' یا کلاس 'irrelevant' تعلق داشته باشد را مشخص می‌کند. به عنوان نمونه، روش مبتنی بر آنتولوژی ممکن است بردار خروجی نظیر {irrelevant: 0.35, relevant: 0.65} را برای یک سند داده شده تولید کند.
- در این مدل، واحدی به نام "Gating Network" وجود دارد. این واحد بر اساس یادگیری انجام شده روی داده‌های ورودی، به هرکدام از خبره‌ها یک وزن مشخص اختصاص می‌دهد [۲۹، ۳۰]. بدین منظور سه مجموعه داده مهیا شده است که هر کدام حاوی ۲۰۰ سند از مجموعه داده 20Newsgroup و Reuters-21578 هستند (این مجموعه‌های داده از داده‌های استفاده شده در فاز یادگیری متفاوت هستند).
- در ابتدا، وزن برابری به هرکدام از خبره‌ها اختصاص داده می‌شود.
- در سه مرحله جداگانه، واحد "Gating Network" عملکرد هر کدام خبره‌ها را می‌سنجد (یاد می‌گیرد). به عبارت دیگر، در هر مرحله عملکرد روش‌های دانش‌محور فیلتر اطلاعات روی یکی از مجموعه‌های داده‌ای تدارک دیده شده سنجیده می‌شود.
- در هر مرحله، خطای خبره‌ها توسط رابطه زیر محاسبه می‌شود.

$$\alpha_i = \frac{1}{2}\log((1-error)/error)\ where\ error = \frac{\sum_{j=1}^{n}|y_{ij}-T_j|^2}{n} \quad (9)$$

در این رابطه، y_{ij} خروجی خبره iام برای سند jام است، T_j خروجی مطلوب برای سند jام، i شاخص هر خبره و n تعداد اسناد ورودی در هر مرحله از ارزیابی عملکرد واحد"Gating Network" است.

- با استفاده از رابطه زیر، خطای محاسبه شده در هر مرحله برای به‌روزرسانی وزن‌های خبرگان استفاده می‌شود.

$$w_i = \frac{w_i * e^{\alpha_i}}{\sum w_i * e^{\alpha_i}} \quad (10)$$

- بعد از مرحله سوم، وزن‌های اختصاص داده شده به هر خبره توسط بردار $weight = \{w_i(e)\}_{i=1}^{k}$ نمایش داده می‌شود.
- پس از وزن‌دهی خبرگان توسط واحد "Gating Network"، فیلتر اسناد آغاز می‌شود. خروجی تمامی خبرگان (دسته‌بندهای معنایی) توسط تابع ترکیب (جمع وزن‌دار خروجی‌ها) تجمیع می‌شود. سپس مقدار حاصل با آستانه β مقایسه می‌شود.

۵- ارزیابی و نتایج

۵-۱- تنظیمات بخش ارزیابی

سیستم فیلتر و بازیابی اطلاعات پیشنهادی در محیط برنامه‌نویسی جاوا توسعه داده شده است. داده‌های ویکی‌پدیا و BNC به صورت

رایگان و توسط اعضای پروژه تحقیقاتی D.I.S.C.O [۲۲] در دسترس عموم قرار گرفته شده است. مجموعه داده ویکی‌پدیا حاوی بیش از ۲ میلیارد توکن و BNC نیز حاوی ۱۲۰ میلیون توکن است. کتابخانه WordNet :: Similarity ([۳۱]) به عنوان مبنایی برای نوشتن کدهای واحد محاسبه معنایی شباهت در زبان برنامه‌نویسی جاوا در نظر گرفته شده است.

برای ارزیابی روش‌های پیشنهادی محاسبه معنایی شباهت، مجموعه داده‌های M-C [۳۲، ۳۳] استفاده شده است. این مجموعه داده حاوی ۳۰ جفت کلمه است که توسط گروهی متشکل از ۳۸ سوژه انسانی رتبه‌بندی شده‌اند؛ سی جفت کلمه در مقیاس صفر (عدم شباهت) تا چهار (کاملاً هم‌معنی) رتبه‌بندی شده‌اند. این مجموعه داده به عنوان معیاری معتبر برای ارزیابی روش‌های محاسبه معنایی شباهت محسوب می‌شود [۳۴]. برای ارزیابی عملکرد سیستم فیلتر محتوایی اطلاعات، مجموعه داده 20Newsgroup [۳۵] و مجموعه داده Reuters-21578 [۳۶] استفاده شده است.

در هر مرحله از ارزیابی، فرض بر این است که ترجیحات کاربری تنها در یکی از کلاس‌ها قابل دسته‌بندی است و کاربر به محتوای اطلاعاتی اسناد موجود در دیگر کلاس‌ها علاقه‌مند نیست. اسناد موجود در مجموعه داده 20Newsgroup را می‌توان در پنج دسته‌بندی موضوعی گسترده‌تر، یعنی "Computer"، "Politics"، "Science"، "Recreation" و "Religion" دسته‌بندی کرد. در این راستا، مجموعه‌ای از پنج آزمایش برای ارزیابی عملکرد سیستم طراحی شده است. در هر آزمایش، اسناد دسته‌بندی شده در یکی از دسته‌بندی‌های موضوعی، به عنوان اسناد «مرتبط» با ترجیحات کاربری و اسناد موجود در دیگر دسته‌بندی‌های موضوعی به عنوان اسناد «نامرتبط» شناخته می‌شوند. همچنین، در مجموعه داده Reuters-21578، پنج کلاس از مجموعه کلاس‌های موجود یعنی "earn"، "Acq"، "interest"، "trade" و "crude" برای ارزیابی عملکرد سیستم پیشنهادی انتخاب شده‌اند. برای ارزیابی سیستم روی مجموعه داده Reuters-21578 نیز پنج آزمایش مانند مجموعه داده قبلی طراحی شده است. از آنجایی که سیستم بر اساس مفهوم دسته‌بندی باینری کار می‌کند، بهترین مقدار برای پارامتر β (بخش ۵-۵) مقدار ۰.۵ است.

۵-۲- ارزیابی همبستگی روش‌های محاسبه معنایی شباهت با قضاوت انسانی

در مرحله اول از ارزیابی ، همبستگی روش‌های محاسبه معنایی شباهت با قضاوت انسانی با استفاده از فرمول همبستگی پیرسون محاسبه می‌شود. همبستگی روش‌های محاسبه معنایی شباهت با مجموعه داده M-C (قضاوت انسانی) در جدول ۱ به تصویر کشیده شده است. نماد (*) نتیجه روش‌های توسعه داده شده در این مقاله را نشان می‌دهد.

جدول ۱- نتایج ارزیابی همبستگی روش‌های پیشنهادی با قضاوت انسانی

روش	همبستگی با قضاوت انسانی
M-C dataset	1.0 (Grand Truth)
Web DICE [37]	0.267
Web Overlap [37]	0.382
Sahami [37]	0.579
CODC [37]	0.693
ESA [37]	0.58
Web PMI [37]	0.548
BNC First-Order (*)	0.617
BNC Second-Order (*)	0.651
Wikipedia First-Order (*)	0.776
Wikipedia Second-Order (*)	0.897
WordNet(*)	0.413 - 0.648

مقایسه نتایج در جدول ۱، نشان می‌دهد که روش‌های مبتنی بر ویکی‌پدیا نسبت به سایر روش‌های پیشنهادی، همبستگی بیشتری با قضاوت انسانی دارند. با توجه به ساختار غنی و معنایی ویکی‌پدیا و برتری اثبات شده آن در مدل‌سازی دانش دقیق و مفید در دامنه‌های مختلف اطلاعاتی، نتایج حاصل از روش‌های مبتنی بر ویکی‌پدیا تا حدودی قابل انتظار بودند. همچنین، با مقایسه نتایج با سایر روش‌های مشابه محاسبه معنایی شباهت، به این نتیجه می‌رسیم که روش‌های مبتنی بر ویکی‌پدیا و BNC بهتر از روش‌های شناخته شده مانند CODC و ESA عمل می‌کنند

۵-۳- ارزیابی چارچوب فیلتر اطلاعات پیشنهادی

در مرحله بعدی، مقدار Recall، Precision، Accuracy و F-Measure روش‌های پیشنهادی روی مجموعه داده 20Newsgroup سنجیده می‌شود. این آزمایش‌ها برای ارزیابی عملکرد و دقت روش‌های پیشنهادی در شناسایی و دسته‌بندی اسناد مرتبط با ترجیحات کاربری طراحی شده‌اند.

مجموعه‌ای از پنج آزمایش روی مجموعه داده 20Newsgrouop طراحی شده است. در هر آزمایش، ۱۰۰۰ سند به صورت تصادفی

انتخاب می‌شوند. تعداد ۴۰۰ سند از ۱۰۰۰ سند موجود در داده‌های تست از کلاسی انتخاب شده‌اند که فرض شده است ترجیحات کاربری با آن تطابق دارد و بقیه از چهار کلاس باقیمانده انتخاب می‌شوند (در مجموع ۵۰۰۰ سند به صورت تصادفی انتخاب شده‌اند). در هر آزمایش ترجیحات کاربری توسط یک نمایه کاربری نشان داده می‌شود. چنین نمایه‌ای سیستم را قادر می‌سازد تا مشخص کند که آیا یک سند خاص به کلاس «مرتبط» تعلق دارد یا کلاس «نامرتبط». برای این منظور، در هر آزمایش، اسناد مرتبط با دسته‌بندی موضوعی «مرتبط» با ترجیحات کاربری (متفاوت از اسناد موجود در مجموعه داده تست) تجزیه و تحلیل و مفاهیم/کلمات متناوب و البته حاوی اطلاعات مفید شناسایی می‌شوند. سپس، لیستی از مفاهیم/کلمات کاندید تشکیل و به خبره دامنه اطلاعاتی ارائه می‌شود. با توجه به نتایج حاصل از شباهت محاسبه شده میان نمایه‌ها، اسناد با برچسب "True Positive"، "True Negative"، "False positive" و "False Negative" برچسب‌گذاری می‌شوند. در نهایت، عملکرد سیستم پیشنهادی با توجه به معیارهای Recall، Precision، Accuracy و F-Measure مورد ارزیابی قرار می‌گیرد. نتایج در جدول‌های ۲ تا ۶ نشان داده شده‌اند.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad Recall = \frac{TP}{TP + FN}$$
$$F - measure = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad Precision = \frac{TP}{TP + FP} \tag{11}$$

جدول ۲- ارزیابی روش‌های پیشنهادی روی "COMPUTER" (20Newsgroup)

روش	Precision	Recall	F-Measure	Accuracy
Lin	85.98%	92%	88.89%	90.8%
Jiang & Conrath	88.09%	92.5%	90.24%	92%
Leacock & Chodorow	88.99%	95%	91.9%	93.3%
Path	84.76%	86.25%	85.5%	88.3%
Wu & Palmer	86.86%	89.25%	88.04%	90.3%
Wiki Second-order	93.3%	97.5%	95.35%	96.2%
Wiki First-order	92.33%	96.25%	94.24%	95.3%
BNC Second-order	93.01%	96.5%	94.72%	95.7%
BNC First-order	91.95%	94.25%	93.09%	94.4%
Ontology	96.8%	98.25%	97.52%	98%

جدول ۳- ارزیابی روش‌های پیشنهادی روی "RELIGION" (20Newsgroup)

روش	Precision	Recall	F-Measure	Accuracy
Lin	85.61%	90.75%	88.1%	90.2%
Jiang & Conrath	86.66%	91%	88.78%	90.8%
Leacock & Chodorow	87.47%	94.25%	90.73%	92.3%
Path	88.30%	86.75%	87.52%	90.1%
Wu & Palmer	89.62%	88.5%	89.06%	91.3%
Wiki Second-order	94.81%	96%	95.40%	96.3%
Wiki First-order	90.89%	94.75%	92.78%	94.1%
BNC Second-order	92.42%	94.5%	93.45%	94.7%
BNC First-order	90.12 %	93.5%	91.78%	93.3%
Ontology	95.56%	96.75%	96.15%	96.9%

جدول ۴- ارزیابی روش‌های پیشنهادی روی "RECREATION" (20Newsgroup)

روش	Precision	Recall	F-Measure	Accuracy
Lin	81.15%	85%	83.03%	86.1%
Jiang & Conrath	80.99%	86.25%	83.54%	86.4.%
Leacock & Chodorow	86.57 %	90.25%	88.37%	90.5%
Path	79.15%	83.5%	81.27%	84.6%
Wu & Palmer	80.53 %	83.75%	82.11%	85.4%
Wiki Second-order	94.18%	93%	93.58%	94.9%
Wiki First-order	86.86%	89.25%	88.04 %	90.3%
BNC Second-order	89.33%	90%	89.66%	91.7%
BNC First-order	84.75%	87.5%	86.10 %	88.7%
Ontology	94.13%	92.25%	93.18%	94.6%

جدول ۵ - ارزیابی روش‌های پیشنهادی روی "SCIENCE" (20Newsgroup)

روش	Precision	Recall	F-Measure	Accuracy
Lin	78.86%	83%	80.88%	84.3%
Jiang & Conrath	78.82%	83.75%	81.21%	84.5%
Leacock & Chodorow	84.96%	89%	86.94%	89.3%
Path	75.76%	81.25%	78.41 %	82.1%
Wu & Palmer	77.54%	82%	79.71%	83.3%
Wiki Second-order	90.57%	91.25%	90.9%	92.7%
Wiki First-order	83.49 %	87.25%	85.33%	88%
BNC Second-order	85.68%	88.25%	86.95 %	89.4%
BNC First-order	81%	85.25%	83.07%	86.1%
Ontology	91.75%	91.75%	91.75%	93.4%

جدول ۶- ارزیابی روش‌های پیشنهادی روی "POLITICS" (20Newsgroup)

روش	Precision	Recall	F-Measure	Accuracy
Lin	83.29%	87.25%	85.23%	87.9%
Jiang & Conrath	83.41%	88%	85.64 %	88.2%
Leacock & Chodorow	87.35%	91.5%	89.38%	91.3%
Path	83.21%	84.25%	83.73%	86.9%
Wu & Palmer	84.2%	85.25%	84.72%	87.7%
Wiki Second-order	92.36%	93.75%	93.05 %	94.4%
Wiki First-order	89.05%	91.5%	90.26%	92.1%
BNC Second-order	91.29 %	91.75%	91.52%	93.2%
BNC First-order	86.99%	90.25%	88.59%4	90.7%
Ontology	95.19%	94%	94.59%	95.7%

همان‌طور که در جدول‌های ۲-۶ مشخص است، روش مبتنی بر آنتولوژی بالاترین عملکرد را در میان روش‌های پیاده‌سازی شده دارد (به جز دسته‌بندی موضوعی "Recreation" که در آن روش بردار کلمه مرتبه دوم ویکی‌پدیا بهترین نتایج را به دست آورده است). این نتایج به‌طور مستقیم به ساختار غنی و معنایی آنتولوژی در سازمان‌دهی مفاهیم در قالب مجموعه مفاهیم هم‌معنی و برقراری روابط معنایی میان آن‌ها مرتبط است. روش‌های مبتنی بر بردار کلمه مرتبه دوم ویکی‌پدیا و BNC به ترتیب در مکان دوم و سوم قرار دارند. ساختار غنی اطلاعاتی ویکی‌پدیا و BNC به سیستم اجازه می‌دهد تا ساختارهای معنایی مشترک و مفاهیم مشابه را در نمایه‌ها شناسایی نماید؛ در نتیجه تطابق میان نمایه‌ها دقیق‌تر محاسبه می‌شود. مقدار Recall و Precision بالا برای یک روش خاص بدین معنی است که این روش می‌تواند اسناد مرتبط با ترجیحات کاربری را بهتر از سایر روش‌ها شناسایی کند. از مقایسه نتایج روش‌های مبتنی بر WordNet و Ontology نتایج جالبی به دست آمده است. از آنجایی که ساختار هر دو پایگاه دانش مبتنی بر WordNet است، روش‌های مبتنی بر WordNet نتایج ضعیف‌تری را نسبت به روش مبتنی بر آنتولوژی از خود نشان داده‌اند. با این حال، پس از بررسی بیشتر به این نتیجه رسیدیم که آنتولوژی OntoWordNet روابط قدرتمندتر و متمایزتری میان مفاهیم در مقایسه با WordNet برقرار می‌کند؛ بنابراین، به شکل بهتری می‌تواند شباهت میان نمایه‌ها را مشخص کند. در مرحله بعد، چارچوب یادگیری ترکیبی «ترکیب خبرگان» ارزیابی می‌شود. نتایج در جدول ۷ نشان داده شده است. در این آزمایش، مجموعه داده آزمایش قبل استفاده شده است.

جدول ۷- ارزیابی مدل پیشنهادی ترکیب خبرگان (20Newsgroup)

Topic	Precision	Recall	F-Measure	Accuracy
Computer	98.26%	99%	98.63%	98.9%
Religion	98.24%	97.5%	97.87%	98.3%
Politics	97.24%	96.75%	96.99%	97.6%
Recreation	96.20%	95.0%	95.6%	96.5%
Science	95.44%	94.25%	94.84%	95.9%
Mean Performance	**97.076%**	**96.5%**	**96.789%**	**97.44%**

مطابق جدول ۷ روش «ترکیب خبرگان» از روش‌های موجود در آزمایش قبلی بهتر عمل کرده است و فرض ما مبتنی بر عملکرد و دقت بهتر سیستم در نتیجه یکپارچگی روش‌های مبتنی بر دانش فیلتر اطلاعات با یک مدل یادگیری ترکیبی را تأیید می‌کند. همچنین، نتایج نشان می‌دهند که بسیاری از اسناد به اشتباه دسته‌بندی شده در آزمایش‌های قبلی به درستی توسط روش «ترکیب خبرگان» دسته‌بندی شده‌اند.

در مرحله بعد، عملکرد سیستم پیشنهادی روی داده‌های Reuters-21578 سنجیده می‌شود. این آزمایش‌ها برای ارزیابی عملکرد و دقت روش‌های پیشنهادی در شناسایی اسناد مرتبط با ترجیحات کاربر طراحی شده‌اند. مجموعه‌ای از پنج آزمایش روی مجموعه داده Reuters-21578 طراحی شده است. رویه ارزیابی سیستم فیلتر اطلاعات پیشنهادی روی مجموعه داده 20Newgroup، جهت ارزیابی سیستم روی مجموعه داده Reuters-21578 نیز استفاده می‌شود. نتایج در جدول‌های ۸ تا ۱۲ نشان داده شده است.

جدول ۸- ارزیابی روش‌های پیشنهادی روی "EARN" (REUTERS-21578)

Method	Precision	Recall	F-Measure	Accuracy
Lin	82.71%	83.75%	83.23%	86.5%
Jiang & Conrath	82.49%	81.25%	81.86%	85.6%
Leacock & Chodorow	84.14%	82.25%	83.19%	86.7%
Path	81.68%	80.25%	80.96%	84.9%
Wu & Palmer	83.38%	81.5%	82.43%	86.1%
Wiki Second-order	93.64%	92%	92.81%	94.3%
Wiki First-order	91.14%	90%	90.57%	92.5%
BNC Second-order	91.44%	90.75%	91.09%	92.9%
BNC First-order	90.27%	90.5.%	90.38%	92.3%
Ontology	95.77%	96.25%	96.01%	97.33%

جدول ۹- ارزیابی روش‌های پیشنهادی روی "ACQ" (REUTERS-21578)

Method	Precision	Recall	F-Measure	Accuracy
Lin	82.71%	82.5%	82.6%	86.1%
Jiang & Conrath	81.16%	80.75%	80.95%	84.8%
Leacock & Chodorow	84.13%	83.5%	83.81%	87.1%
Path	77.0%	79.5%	78.23%	82.3%
Wu & Palmer	82.7%	81.25%	81.97%	85.7%
Wiki Second-order	93.13%	91.5%	92.31%	93.9%
Wiki First-order	89.08%	89.75%	89.41 %	91.5%
BNC Second-order	89.55%	90 %	89.78%	91.8%
BNC First-order	86.89%	89.5%	88.18%	90.4%
Ontology	95.99%	95.75%	95.87 %	96.7 %

جدول ۱۰- ارزیابی روش‌های پیشنهادی روی "INTEREST" (REUTERS-21578)

Method	Precision	Recall	F-Measure	Accuracy
Lin	79.51%	81.5 %	80.49%	84.2%
Jiang & Conrath	79.16 %	79.75%	79.75 %	83.5%
Leacock & Chodorow	83.59 %	82.75%	83.17%	86.6%
Path	73.62%	76.75%	75.15%	79.7 %
Wu & Palmer	81.11 %	80.5%	80.8 %	84.7%
Wiki Second-order	91.13%	92.5%	91.81 %	93.4%
Wiki First-order	87.91 %	87.25%	87.58 %	90.1%
BNC Second-order	87.78 %	88 %	87.89 %	90.3%
BNC First-order	83.54 %	86.25 %	84.87 %	87.7%
Ontology	95.93 %	94.25%	95.08 %	96.1%

جدول ۱۱- ارزیابی روش‌های پیشنهادی روی "TRADE" (REUTERS-21578)

Method	Precision	Recall	F-Measure	Accuracy
Lin	82.04%	82.25%	82.15%	85.7%
Jiang & Conrath	79.95 %	80.75%	80.35%	84.2%
Leacock & Chodorow	83.58 %	84%	83.79%	87%
Path	76.53%	78.25%	77.38%	81.7 %
Wu & Palmer	82.8%	81.75%	82.26 %	85.9%
Wiki Second-order	91.38%	92.75%	92.06%	93.6%
Wiki First-order	88.89%	88%	88.44 %	90.8%
BNC Second-order	89.25%	89.25 %	89.25%	91.4%
BNC First-order	84.8 %	89.25%	86.97%	89.3%
Ontology	98.19 %	95%	96.57 %	97.3 %

جدول ۱۲- ارزیابی روش‌های پیشنهادی روی "CRUDE" (REUTERS-21578)

Method	Precision	Recall	F-Measure	Accuracy
Lin	76.17%	77.5 %	76.83%	81.3%
Jiang & Conrath	76.88 %	76.5%	76.69 %	81.4%
Leacock & Chodorow	79.27 %	81.25%	80.25 %	84%
Path	67.12%	74.5%	70.62%	75.2 %
Wu & Palmer	76.46 %	78.75%	78.59%	81.8%
Wiki Second-order	88.72%	90.5%	89.6 %	91.6%
Wiki First-order	85.51 %	85%	85.75 %	88.7%
BNC Second-order	85.0 %	89.25%	87.07%	89.4%
BNC First-order	83.25 %	84.5%	83.87 %	87%
Ontology	94.21%	93.5%	93.85 %	95.1%

همان‌طور که در جدول‌های ۸ تا ۱۲ قابل مشاهده است، روش مبتنی بر آنتولوژی از سایر روش‌های پیاده‌سازی شده بهتر عمل می‌کند. این روش در تمامی کلاس‌های مجموعه داده Reuters-21578 بهترین عملکرد را از خود نشان می‌دهد. همچنین روش مبتنی بر بردار کلمه مرتبه دوم ویکی‌پدیا نتایج خوبی از خود نشان می‌دهد؛ به‌ویژه در مقایسه با روش‌های مبتنی بر WordNet و BNC. همان‌طور که از نتایج مشهود است، روش‌های مبتنی بر WordNet نتایج ضعیفی در مقایسه با دیگر روش‌های فیلتر محتوایی اطلاعات به دست آورده‌اند. دلیل نتایج ضعیف این است که اسناد موجود در مجموعه داده Reuters-21578 در دسته‌بندی‌های موضوعی با محتوای اطلاعاتی بسیار متمایز از هم سازمان‌دهی می‌شوند (بر خلاف مجموعه داده 20Newsgroup که در پنج دسته‌بندی موضوعی با محتوای اطلاعاتی مشابه دسته‌بندی می‌شوند). بنابراین، اگرچه پایگاه دانش WordNet و BNC از ساختار اطلاعاتی غنی بهره می‌برند، اما محتوای اطلاعاتی و روابط معنایی برقرار شده میان موجودیت‌های اطلاعاتی آن‌ها به اندازه کافی جامع نیست تا همه دسته‌بندی‌های موضوعی مربوط به مجموعه داده Reuters-21578 را پوشش دهد. در مرحله بعد، چارچوب یادگیری ترکیبی «ترکیب خبرگان» با توجه به معیارهای Recall، Precision، Accuracy و F-Measure ارزیابی می‌شود.

جدول ۱۳- ارزیابی مدل پیشنهادی ترکیب خبرگان (REUTERS-21578)

Category	Precision	Recall	F-Measure	Accuracy
Earn	98.25 %	98.25%	98.25 %	98.6%
Acq	97.26%	97.5%	97.38%	97.9%
Trade	97.23%	96.5%	96.86%	97.5%
Interest	96.73%	96%	96.36%	97.1%
crude	95.73%	95.25%	95.49%	96.4%
Mean *Performance*	**97.04%**	**96.7%**	**96.863%**	**97.5%**

مطابق با نتایج جدول ۱۳، روش «ترکیب خبرگان» از سایر روش‌ها بهتر عمل می‌کند. همچنین روش «ترکیب خبرگان»، در تمامی کلاس‌های مجموعه داده Reuters-21578 عملکرد بالایی از خود نشان می‌دهد. نتایج بار دیگر فرض ما را برتر بودن روش پیشنهادی «ترکیب خبرگان» در فیلتر و مدیریت اطلاعات تأیید می‌کند.

در ادامه، تعدادی روش یادگیری شناخته شده برای دسته‌بندی موضوعی اسناد پیاده‌سازی شده‌اند که روی مجموعه داده‌های 20Newsgroup و Reuters-21578 ارزیابی می‌شوند. نتایج ارزیابی این روش‌ها با روش «ترکیب خبرگان» پیشنهادی مقایسه می‌شود. سه روش یادگیری ماشین برای دسته‌بندی موضوعی اسناد پیاده‌سازی شده‌اند: (۱) Extreme gradient boosting [۳۸]، (۲) Random Forest (Bagging) [۳۹] و (۳) Recurrent Neural Network-LSTM (LSTM Network) [۴۰].

جدول ۱۴ ویژگی‌های قابل استخراج از مجموعه داده‌ها را نشان می‌دهد. علاوه بر ویژگی‌های استاندارد، شبکه عصبی LSTM با استفاده از ویژگی‌های Word_Embeddings آموزش داده شده است. جدول ۱۵ خصوصیات مدل‌های یادگیری را نشان می‌دهد. عملکرد سیستم با پارامترهای مختلف سنجیده و بهترین پارامترها برای ارزیابی نهایی انتخاب شده‌اند.

جدول ۱۴- ویژگی‌های قابل استخراج از مجموعه داده‌ها

Features	Description	Usage
TF-IDF Feature	Word Level TF-IDF N-gram Level TF-IDF (N=2)	training the Extreme gradient Boosting and Random Forest algorithms
Text-Based Semantic Features	Noun Count, Verb Co Adjective Count, Adverb Count, Pronoun Count	training the Extreme gradient Boosting and Random Forest algorithms
Topic Models Features	Latent Dirichlet Allocation (LDA) for generating Topic Modelling Features,	training the Extreme gradient Boosting and Random Forest algorithms
Word_Embeddings	Word_embeddings.	Training LSTM Deep Neural Network.

جدول ۱۵- خصوصیات مدل‌های یادگیری

Method	Features	Properties
XGB	TF-IDF Feature; Text-Based Semantic Features; Topic Models Features;	Sample_size (20NG)=18820 ; (Reuters)=13328; Test/train _split=35/65;
Random Forest	TF-IDF Feature; Text-Based Semantic Features; Topic Models Features;	Sample_size (20NG)=18820 ; (Reuters)=13328; Test/train _split=35/65;
RNN-LSTM	Word_Embeddings (Word2Vec)	Sample_size (20NG)=18820 ; (Reuters)=13328; Test/train _split=35/65;

عملکرد روش‌های یادگیری ماشین روی مجموعه داده 20Newsgroup در جدول ۱۶ نشان می‌دهد که روش‌های Random Forrest و Extreme Gradient Boosting بهترین نتایج را در بین روش‌های یادگیری ماشین دارند. با این حال، شبکه عصبی LSTM نتایج ناامیدکننده‌ای در مقایسه با دیگر روش‌ها روی مجموعه داده 20Newsgroup کسب کرده است.

جدول ۱۶- عملکرد روش‌های یادگیری ماشین روی مجموعه داده 20Newsgroup

Methods	Mean Accuracy	Mean Precision	Mean Recall	Mean F-measure
XGB	***96.745%***	***92.538%***	***92.287%***	***92.403%***
Random Forest	***97.131%***	***93.729%***	***93.542%***	***93.325%***
RNN-LSTM	***93.308%***	***90.853%***	***92.055%***	***91.448%***

عملکرد روش‌های یادگیری ماشین روی مجموعه داده Reuters-21578 در جدول ۱۷ نشان می‌دهد که روش‌های Random Forest و شبکه عصبی LSTM بهترین نتایج را در میان روش‌های مبتنی بر یادگیری ماشین کسب کرده‌اند. با این حال و به نظر ما، شبکه عصبی LSTM نتایج بهتر و قابل‌اعتمادتری نسبت به دیگر روش‌های یادگیری ماشین در کلاس‌های مجموعه داده Reuters-21578 کسب کرده است و مقادیر Precision بهتری از خود نشان می‌دهد

نتایج نشان می‌دهند که ساختار یادگیری ترکیبی «ترکیب خبرگان» در مدل‌سازی معنایی محتوای اطلاعاتی اسناد و همچنین فیلتر اطلاعات مؤثرتر عمل می‌کند. نتایج ناامیدکننده شبکه عصبی LSTM را نیز می‌توان به تبدیل عددی ویژگی‌های معنایی موجود در محتوا در فرایند یادگیری مرتبط دانست. در نتیجه چنین تبدیلی بخشی از معنا و محتوای اطلاعاتی موجود در اسناد از بین می‌رود.

جدول ۱۷- عملکرد روش‌های یادگیری ماشین روی مجموعه داده Reuters-21578

Methods	Mean Accuracy	Mean Precision	Mean Recall	Mean F-measure
XGB	***94.177%***	***91.879%***	***91.723%***	***91.801%***
Random Forest	***94.517%***	***91.727%***	***92.963%***	***92.34%***
RNN-LSTM	***94.432%***	***92.152%***	***92.189%***	***92.169%***

در ادامه، دقت (Accuracy) برخی از روش‌های شناخته شده یادگیری عمیق و یادگیری ماشین با سیستم ترکیبی (یادگیری- دانش‌محور) پیشنهادی مورد مقایسه قرار می‌گیرد. نتایج ارزیابی و مقایسه انجام شده در جدول‌های ۱۸ و ۱۹ نمایش داده شده‌اند.

جدول ۱۸- ارزیابی روش‌های شناخته شده یادگیری عمیق و یادگیری ماشین با سیستم ترکیبی پیشنهادی (20Newsgroup)

روش	ساختار استفاده شده	دقت (Accuracy)
Jiang et al [41]	DBN+Softmax	85.57%
.Shih et al [42]	LSTM	86.2%
Camacho [43]	CNN, LSTM	90.9%
Wang et al [44]	Cross domain Transfer learning	95.62%
Asim et al [45]	Hybrid Learning	91.729%
روش پیشنهادی	Hybrid (KB-Learning)	97.44%

جدول ۱۹- ارزیابی روش‌های شناخته شده یادگیری عمیق و یادگیری ماشین با سیستم ترکیبی پیشنهادی (Reuters-21578))

روش	ساختار استفاده شده	دقت (Accuracy)
Kowsari et al [46]	Deep Learning	90.96%
Revanasiddappa et al [47]	Symbolic cluster based	86.75%
روش پیشنهادی	Hybrid (KB-Learning)	97.5%

نتایج نمایش داده شده نشان می‌دهد که روش ترکیبی پیشنهادی در این مقاله از روش‌های شناخته شده یادگیری ماشین و یادگیری عمیق بهتر عمل می‌کند؛ به عبارت دیگر، اسناد متنی کمتری توسط این روش پیشنهادی به شکل نادرست دسته‌بندی می‌شوند.

۶- بحث و نتیجه‌گیری

در این مقاله، یک روش ترکیبی (یادگیری ماشین-دانش‌محور) برای فیلتر و مدیریت اطلاعات معرفی شده است. سیستم پیشنهادی شامل چندین واحد محاسبه معنایی شباهت مبتنی بر دانش است که در سیستم فیلتر اطلاعات یکپارچه شده‌اند. این واحدها از مفهوم شباهت معنایی میان مفاهیم برای مشخص کردن میزان شباهت سند ورودی به ترجیحات کاربری استفاده می‌کنند. یکی از مزایای اصلی سیستم ترکیبی پیشنهاد شده وابستگی کم آن‌ها به فرایند یادگیری است. بنابراین، هرگونه تغییر در دامنه اطلاعاتی سیستم منجر به تنزل عملکرد فرایند فیلتر اطلاعات نمی‌شود. تنها یادگیری زمانی اتفاق می‌افتد که "Gatin Network" در حال یادگیری و اختصاص وزن به واحدهای شباهت معنایی است. همچنین از طریق یکپارچه‌سازی دانش ساخت‌یافته آنتولوژی و پایگاه‌های دانش در تمامی واحدهای سیستم فیلتر اطلاعات، مسئله ابهام محتوا در کاربردهای متن‌کاوی را مورد بررسی قرار داده‌ایم. ارزیابی روش‌های محاسبه معنایی شباهت همبستگی بالای این روش‌ها با قضاوت انسانی و در نتیجه مطابقت دقیق‌تر میان نمایه‌ها را نشان می‌دهد همچنین نتایج حکایت از آن دارند که پایگاه‌های دانش مانند ویکی‌پدیا حاوی اطلاعات دقیق در مورد دامنه‌های اطلاعاتی مختلف هستند و می‌توانند به عنوان ابزاری مفید در تمامی واحدهای یک سیستم فیلتر و مدیریت اطلاعات استفاده شوند. همچنین، ارزیابی سیستم ترکیبی پیشنهادی روی مجموعه داده‌های 20Newsgroup و Reuters-21578 حکایت از برتری سیستم پیشنهادی بر سیستم‌های شناخته شده مبتنی بر یادگیری ماشین دارند. با توجه به نتایج حاصل شده می‌توان نتیجه گرفت که آنتولوژی و ویکی‌پدیا ساختار اطلاعاتی غنی‌تر نسبت به BNC و WordNet دارند و برای کاربردهای فیلتر و مدیریت اطلاعات مناسب‌تر هستند. در نهایت، ارزیابی نشان می‌دهد که ترکیب خبره (تخصص) جمعی/گروهی روش‌های دانش‌محور با یک مدل یادگیری نقش مؤثری در افزایش دقت و بهبود عملکرد سیستم خواهد داشت. هر کدام از روش‌های مبتنی بر دانش می‌توانند به‌طور مستقل برای دسته‌بندی معنایی محتوا استفاده شوند. با این حال، ترکیب آن‌ها تأثیر مستقیمی در افزایش دقت و عملکرد خواهد داشت. همان‌طور که در این مقاله نشان داده شده است، سیستم پیشنهادی می‌تواند به عنوان یک دسته‌بند معنایی، تجزیه و تحلیل معنایی جامعی از اسناد ارائه دهد و مناسب‌ترین دسته‌بندی موضوعی برای اسناد پیشنهاد دهد. به‌عنوان کارهای آینده، می‌توانیم روش‌های یادگیری عمیق را در برخی از جنبه‌های سیستم فیلتر اطلاعات پیشنهادی (به عنوان مثال استخراج ویژگی) یکپارچه نماییم و تأثیرات آن را بررسی کنیم.

۷- مراجع

[1] P. Shoval, V. Maidel, B. Shapira, "An Ontology- Content-Based Filtering Method", International Journal "Information Theories & Applications", Vol. 15, pp.303-314, 2008.

[2] K. N. Junejo, A. Karim, M. T. Hassan, M. Jeon, "Terms-based discriminative information space for robust text classification", Information Sciences, Vol. 372, pp.518-538, 2016.

[3] X. Zhang, X. Hou, X. Chen, T. Zhuang, "Ontology-based semantic retrieval for engineering domain knowledge", Neurocomputing, Vol. 116, pp.382-391, 2013.

[4] Y. Jiang, W. Bai, X. Zhang, J. Hu, "Wikipedia-based information content and semantic similarity computation", Information Processing & Management, Vol. 52, pp.248-265, 2017.

[5] B. Shapira, N. Ofek, and V. Makarenkov, "Exploiting Wikipedia for Information Retrieval Tasks", In Proceedings of 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.1137-1140, 2015.

[6] O. Medelyan, D. Milne, C. Legg and I. Witten, "Mining meaning from Wikipedia", International Journal of Human-Computer Studies, Vol. 67, pp.716–754, 2009.

[7] H. K. Kim, H. Kim, and S. Cho, "Bag-of-concepts: Comprehending document representation through clustering words in distributed representation", Neurocomputing, Vol. 266, pp.336-352, 2017.

[8] Y. Shen, X. He, J. Gao, L. Deng, and G. Mesnil, "A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval", In Proceedings of *CIKM*, pp.101–110, 2014.

[9] H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song And R. Ward, "Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval", IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 24, pp.694-707, 2016.

[10] V. Maidel, P. Shoval, B. Shapira, M. Taieb-Maimon, "Ontological content-based filtering for personalized newspapers: A method and its evaluation", Online Information Review, Vol. 34, pp.729 – 756, 2010.

[11] S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Cicekli, F.N. Alpaslan, "An ontology-based retrieval system using semantic indexing", Information Systems, Vol. 37, pp.294-305, 2012.

[12] G-J. Hahm, J-H. Lee, H-W. Suh, "Semantic relation based personalized ranking approach for engineering document retrieval", Advanced Engineering Informatics, Vol. 29, pp.366-379, 2015.

[13] M. Daoud, L. Tamine, M. Boughanem, "A personalized search using a semantic distance measure in a graph-based ranking model", Journal of Information Science, Vol. 37, pp.614–636, 2011.

[14] M. A. H. Taieb, M. B. Aouicha, A. B. Hamadou,"Computing semantic relatedness using Wikipedia features", Knowledge-Based Systems", Vol. 50, pp.260-278, 2013.

[15] P Malo, A. Sinha, J. Wallenius and P. Korhonen, "Concept-based Document Classification Using Wikipedia and Value Function", Journal of the American Society for Information Science and Technology, Vol. 62, pp.2496–2511, 2011.

[16] E. Gabrilovich and S. Markovitch, "Computing semantic relatedness using Wikipedia-based explicit semantic analysis", In Proceedings of the 20th international joint conference on Artifical intelligence,

the 16th international conference on World Wide Web, pp.757-766, 2007.

[38] R. Song, S. Chen, B. Deng, and L. Li, "eXtreme Gradient Boosting for Identifying Individual Users Across Different Digital Devices", In Proceedings of WAIM, Vol. 9658, pp. 43–54, 2016.

[39] Q. Wu, Y. Ye, H. Zhang, M. Ng and S. Ho, "ForesTexter: An efficient random forest algorithm for imbalanced text categorization", Knowledge-Based Systems, Vol. 67, pp.105-116, 2014.

[40] G. Rao, W. Huang, Z. Feng and Q. Cong, "LSTM with sentence representations for document-level sentiment classification", Neurocomputing, Vol. 308, pp.49-57, 2018.

[41] M. Jiang, Y. Liang, X. Feng, X. Fan, Z. Pei, Y. Xue, and R. Guan, "Text classification based on deep belief network and softmax regression," Neural Computing and Applications, vol. 29, no. 1, pp. 61–70, 2018.

[42] C.-H. Shih, B.-C. Yan, S.-H. Liu, and B. Chen, "Investigating siamese lstm networks for text categorization," in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2017, pp. 641–646.

[43] J. Camacho-Collados and M. T. Pilehvar, "On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis," arXiv preprint arXiv:1707.01780, 2017.

[44] D. Wang, C. Lu, J. Wu, H. Liu, W. Zhang, F. Zhuang, and H. Zhang, "Softly associative transfer learning for cross-domain classification," IEEE transactions on cybernetics, 2019.

[45] , M. N. Asim, M. U. G. Khan, M. I. Malik, A. Dengel, S. Ahmed, "A Robust Hybrid Approach for Textual Document Classification"., arXiv preprint arXiv:1909.05478, 2019.

[46] K. Kowsari, M. Heidarysafa, D. E. Brown, K. J. Meimandi, L. Barnes, "Rmdl: Random multimodel deep learning for classification", In Proceedings of the 2nd International Conference on Information System and Data Mining, pp. 19-28, 2018.

[47] M. B. Revanasiddappa, B. S. Harish, S. Manjunath, "Document classification using symbolic classifiers", In 2014 International Conference on Contemporary Computing and Informatics, pp. 299-303, 2014.

pp.1606-1611, 2007.

[17] R. Navigli and S. P. Ponzetto, "Babelrelate! A Joint Multilingual Approach to Computing Semantic Relatedness", In Proceedings of the 26th AAAI Conference on Artificial Intelligence, pp.108-114, 2012.

[18] Z. Wu, H. Zhu, G. Li, Z. Cui, H. Huang, J. Li, E. Chen, G. Xu, "An efficient Wikipedia semantic matching approach to text document classification", Information Sciences, Vol. 393, 15-28, 2017.

[19] J. Gao, B. Zhang, X. Chen, "A WordNet-based semantic similarity measurement combining edge-counting and information content theory", Engineering Applications of Artificial Intelligence, Vol. 39, pp.80-88, 2015.

[20] OntoWordNet Ontology, "Laboratory for applied ontology - DOLCE", [last visited on Feb. 19, 2013], [Online], Available: <http://www.loa.istc.cnr.it/DOLCE.html# OntoWordNet >.

[21] J. G. Mersch, R. R. Lang, "An Information-Theoretic Sentence Similarity Metric", In Proceedings of the 28th International Florida Artificial Intelligence Research Society Conference, pp.552-556, 2015.

[22] P. Kolb, "DISCO: A Multilingual Database of Distributionally Similar Words", In Proceedings of KONVENS, pp.5-12, 2008.

[23] M. Warin, "Using WordNet and Semantic Similarity to Disambiguate an Ontology", Thesis for the degree of Doctor of Philosophy, STOCKHOLMS University, Institute of Linguistic, 2004.

[24] C. Biemann, S. P. Ponzetto, S. Faralli, A. Panchenko, and E. Ruppert, "Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation.", In Proceedings of EACL, 2017.

[25] W. Cohen, P. Ravikumar, S. Fienberg, "A comparison of string distance metrics for name-matching tasks", In Proceedings of International Conference on Information Integration on the Web, pp.73-78, 2003.

[26] R. Thiagarajan, G. Manjunath and M. Stumptner, "Computing Semantic Similarity Using Ontologies", HP Laboratories, 2008.

[27] L. Luna, R. Quintero, M. Torres, M. Moreno-Ibarra, G. Guzmán, I. Escamilla, "An ontology-based approach for representing the interaction process between user profile and its context for collaborative learning environments", Computers in Human Behavior, Vol. 51, pp.1387-1394, 2015.

[28] D. Lin, "Extracting Collocations from Text Corpora", In Workshop on Computational Terminology, pp.57–63, 1998.

[29] L. Xu, S. Amar, "Combining Classifiers and Learning Mixture-of-Experts", Encyclopaedia of Artificial Intelligence, Vol. 3, pp.318-326, 2009.

[30] S. Masoudnia and and R. Ebrahimpour, "Mixture of experts: a literature survey", Artificial Intelligence Review, Vol. 42, pp.275-293, 2014.

[31] Ted Pedesen, "WordNet::Similarity", [last visited on Sep. 29, 2016], [Online] Available: <http://wn-similarity.sourceforge.net/>.

[32] G.A. Miller and W.G. Charles, "Contextual correlates of semantic similarity". Language and Cognitive Processes, Vol. 6, pp.1-28, 1991.

[33] D. Girardi, S. Wartner, G. Halmerbauer, M. Ehrenmüller, H. Kosorus, S. Dreiseitl, "Using concept hierarchies to improve calculation of patient similarity", Journal of Biomedical Informatics, Vol. 63, pp.66-73, 2016.

[34] Mehmet Ali Salahli, "An Approach for Measuring Semantic Relatedness between Words via Related Terms", Mathematical and Computational Applications, Vol. 14, pp.55-63, 2009.

[35] K. Lang, "The 20Newsgroups data set, version 20news-18828", [last visited on Sep. 29, 2016], [Online] Available: <http://www.qwone.com/~jason/20Newsgroups>.

[36] W. Zhang, X. Tang, T. Yoshida, "TESC: An approach to TExt classification using Semi-supervised Clustering", Knowledge-Based Systems, Vol. 75, pp.152-160, 2015.

[37] D. Bollegala, Y. Matsuo and M. Ishizuka, "Measuring semantic similarity between words using web search engines", In Proceedings of