توصیهٔ برچسب در شبکه‌های اجتماعی به‌کمک خلاصه‌سازی متن و k-نزدیک‌ترین همسایه

مهسا رحیمی رسکتی[1]*، همایون موتمنی[2]، ابراهیم اکبری[3]، حسین نعمت‌زاده[4]

استادیار معلم، آموزش و پرورش استان مازندران، ساری، ایران[1]*

استاد (پروفسور) دانشکدهٔ مهندسی کامپیوتر، دانشگاه آزاد ساری، مازندران، ساری، ایران [2]

دانشیار دانشکدهٔ مهندسی کامپیوتر، دانشگاه آزاد ساری، مازندران، ساری، ایران[3]

استادیار دانشکدهٔ مهندسی کامپیوتر، دانشگاه آزاد ساری، مازندران، ساری، ایران[4]

چکیده

امروزه استفاده از شبکه‌های اجتماعی به میزان فزاینده‌ای افزایش یافته‌است. یکی از مهم‌ترین مسائل در این خصوص بالابردن میزان بازدید پست یا پیام افراد است و بهترین عاملی که می‌تواند در این امر کمک کند، برچسب‌ها[1] هستند. برچسب‌ها به‌طور گسترده‌ای در نظم‌دهی و جست‌وجو در میان داده‌های عظیم موجود نقش دارند، در حال حاضر ایجاد برچسب به‌صورت خودکار به‌شدت مورد توجه قرار گرفته‌است. در این مقاله سعی شده‌است تا به‌کمک خلاصه‌سازی متن، از روی داده‌ها، یک‌سری واژه‌های کلیدی پیشنهادی ایجاد کرد و به‌کمک آن یک پیشنهاددهندهٔ برچسب ارائه کرد؛ بنابراین این مقاله با ترکیب روش‌های خوشه‌بندی، خلاصه‌سازی و توصیهٔ پیشنهاد راه‌کار جدیدی ارائه داده‌است. در روش پیشنهادی به‌کمک مدل مخزن واژه‌های (BoW[2])، تحلیل معنایی آشکار (ESA[3]) و ترکیب آن‌ها با الگوریتم k-نزدیک‌ترین همسایه (K-NN)، یک‌سری برچسب پیشنهادی برای شبکه‌های اجتماعی تهیه شده‌است؛ به این صورت که ابتدا به‌کمک مدل مخزن واژه‌ها، لغت‌نامه‌ای از واژه‌ها ایجاد، سپس به‌کمک الگوریتم k-نزدیک‌ترین همسایه ترکیب‌شده با ESA خوشه‌بندی صحیح و قوی از واژه‌ها به‌وجود می‌آید و درنهایت منجر به برچسب‌های پیشنهادی مناسب می‌شود. راه‌کار پیشنهادی بر روی دو مجموعه‌داده عمومی مورد بررسی قرار گرفت و نتایج برتری خود را نسبت به سایر روش‌های مشابه نشان داد.

واژگان کلیدی: توصیهٔ برچسب، خلاصه‌سازی متن، تعبیهٔ واژه، k نزدیک‌ترین همسایه، BoW.

Tag recommendation in social networks with the help of text summarization and KNN

Mahsa Rahimi Resketi[1*], Homayun Motameni[2],
Ebrahim Akbari [3], Hossein Nematzadeh[4]

Assistant professor & Teacher Education department of Mazandaran, Iran[*1]
professor of Faculty of Computer Engineering, Islamic Azad University, Sari, Mazandaran, Iran[2]
Associate Professor of Faculty of Computer Engineering, Islamic Azad University, Sari, Mazandaran, Iran [3]
Assistant professor of Faculty of Computer Engineering, Islamic Azad University, Sari, Mazandaran, Iran [4]

Abstract

In recent years, the utilization of social networks has surged markedly, with interest in their use escalating daily. A pivotal concern is augmenting the number of views for individuals' posts or messages to enhance their popularity. The most effective means to achieve this objective is through the use of tags. Tags significantly contribute to the organization and retrieval of existing data, and the automatic generation of tags has garnered substantial attention. Tag recommendation from textual sources can be approached as a text extraction issue. This paper endeavors to propose a comprehensive set of suggested

[1] Tag
[2] Bag Of Words
[3] Explicit Semantic Analysis

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۱/۴/۲۸ • تاریخ پذیرش: ۱۴۰۳/۹/۱۴ • تاریخ انتشار: ۱۴۰۳/۱۲/۳۰ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.21.4.15]

keywords derived from data via advanced text summarization techniques, culminating in the presentation of a sophisticated tag recommender. Consequently, this research introduces an innovative and robust solution by integrating clustering, summarization, and recommendation methodologies. Initially, utilizing the Bag of Words (BoW) model, comprehensive word parsing and extraction of word roots are performed. This process yields a bag of words capable of facilitating deep semantic exploration. The data is meticulously simplified to its core elements, with prepositions and repetitions omitted. Verbs, due to their high frequency and significance depending on the context of the sentence or post, are mined separately. Other words are judiciously selected based on their frequency and importance, and stored with their repetition counts. Subsequently, employing the K-Nearest Neighbor (KNN) clustering algorithm, the data is clustered, and the cluster representatives serve as the output tags. A slight modification is made to the KNN algorithm by incorporating the Explicit Semantic Analysis (ESA) method for precise scale calculations.

The proposed solution was rigorously evaluated on two public datasets: TPA, extracted by Aminer, and AG, extracted by ComeToMyHead. The AG dataset comprises 127,600 news articles, categorized into four distinct tag types. Each category contains 30,000 training samples and 1,900 test samples, with a total of 31,900 tags representing global, sports, business, and scientific concepts. The findings of this study were compared with those from 13 similar research papers, which fall into four distinct categories: machine learning, long-short-term memory (LSTM), convolutional neural network (CNN), and capsule-based models. The comparative analysis revealed that the proposed method demonstrates superior accuracy, comprehensive coverage, and an enhanced F-measure.

The integration of advanced text analytics techniques underscores the significance of this study in the broader context of information retrieval and data mining. By harnessing the power of semantic analysis and machine learning, this research provides a novel framework that not only enhances the efficiency of tag recommendation systems but also contributes to the theoretical foundation of automated keyword extraction. The implications of these findings are far-reaching, with potential applications extending beyond social networks to other domains requiring efficient data organization and retrieval.

Keywords: label recommendation, text summarization, word embedding, k-nearest neighbor, BoW.

۱-مقدمه

به دلیل اینکه شبکهٔ اجتماعی محبوبیت فزاینده‌ای در دنیای امروز به‌دست آورده‌است کاربران اینترنت، شروع به استفاده از برچسب‌های مختلف برای علامت‌گذاری و مدیریت منابع وب مانند متن، موسیقی یا تصویر کرده‌اند[۱]. StackOverFlow[1] و Flickr[2] برخی از محبوب‌ترین برنامه‌های وب هستند که به استفادهٔ گسترده از برچسب‌گذاری می‌پردازند. نحوهٔ تولید برچسب‌های مرتبط با محتوا به‌طور خودکار برای هر منبع، موضوع داغ پژوهش‌های امروز است. فرایند برچسب‌گذاری زمان‌بر است؛ بااین‌حال، کیفیت برچسب توصیه‌شده هنوز تا رسیدن به حد مطلوب فاصلهٔ زیادی دارد؛ زیرا وضعیت عملکرد استخراج عبارت کلیدی نسبت به بسیاری دیگر از پردازش‌های زبان طبیعی (NLP) پیچیده و سخت‌تر است[۲].

امروزه چند رسانه نقش اساسی در زندگی بشر ایفا می‌کند. چند رسانه ترکیبی از محتویات مختلف مانند متن، تصویر، صوت، ویدئو، گرافیک و... است که از طریق آن می‌توان به هر نوع اطلاعاتی به‌صورت دیجیتالی دسترسی داشت؛ به عبارت دیگر، ارائهٔ جذابی از داده‌های یک‌پارچه است[۳]. داده‌ها مجموعه‌ای از متغیرها و واحدهای اطلاعاتی‌اند که در یک فرایند جمع‌آوری می‌شوند. داده‌ها را می‌توان به هر شکل و برای هر هدفی مورد استفاده قرار داد یا به آنها دسترسی داشت. فرایندی که برای استخراج داده‌های قابل استفاده از مجموعهٔ قابل توجهی از داده‌های خام استفاده می‌شود، داده‌کاوی نامیده می‌شود. زندگی امروزی داده‌های زیادی ارائه می‌دهد. یافتن اطلاعات مفید از آن‌ها می‌تواند در جنبه‌های مختلف پزشکی، اقتصادی، آموزشی و غیره به ما کمک کند؛ داده‌های متنی یکی از محبوب‌ترین انواع داده‌ها است.

با توجه به میزان بسیار بالای داده‌های متنی، به روشی برای جمع‌آوری اطلاعات خوب و دقیق نیاز است؛ برای مثال در اینترنت این اطلاعات شامل اطلاعات مشتری‌ها، اخبار تلویزیونی، بلاگ‌ها و... است[۴]. خلاصه‌سازی متن، تجریدی[3] است که از محتوای یک یا چند متن به‌دست‌آمده، و حاصل مهم‌ترین اطلاعات متن اصلی بوده‌است، اما به‌طور معمول اندازه‌ای برابر نصف یا کمتر متن اصلی دارد[۵]. خلاصه‌سازی، علم یافتن اطلاعات مهم متن و ارائهٔ آن‌هاست؛ بنابراین، برای یافتن ویژگی‌های مهم، نحوهٔ ارائه و نمایش اهمیت بسیار دارد. یکی از مهم‌ترین نکات در کاوش داده‌ها توجه به ویژگی‌ها[4] و شناسایی آن‌هاست. هر داده‌ای یک‌سری ویژگی خاص دارد که از سایر ویژگی‌ها متمایز بوده و به کاوش

[1] http://stackoverflow.com/tags
[2] https://www.flickr.com/
[3] Abstraction
[4] Feature

[DOI: 10.61186/jsdp.21.4.15]

صحیح کمک می‌کنند. متن، ویژگی‌های مختلف زبانی، معنایی، نحوی، ساختاری و غیره دارد که شاید پرکاربردترین آن‌ها شامل موارد زیر است[۶-۱۰]:

- فرکانس واژه و معکوس فرکانس سند یا (TF/IDF)
- تطابق عنوان متن
- مکان جمله
- طول جمله
- واژه‌های نشانه
- شناسایی بر حسب اسم و فعل
- نشانه‌گذاری‌هایی همچون برچسب‌زنی بخشی از گفت‌وگو یا POS [1] و...

به‌طور معمول خلاصه‌سازی متن در سه گام انجام می‌شودد[۱۱]:

۱. **پیش‌پردازش**: می‌تواند با استفاده ابزار کمک فرایند پردازش زبان طبیعی[2] و یا ابزار زبان طبیعی پایتون[3] انجام گیرد.

۲. **امتیاز میزان اطلاعاتی که یک جمله در بردارد**: بر طبق فاکتورهای گوناگون میزان بااهمیت‌بودن و میزان اطلاعاتی که یک متن دارد محاسبه می‌شود تا در ادامه در جهت حذف و یا عدم حذف یک جمله به‌کار رود.

۳. **استخراج و ایجاد خلاصه**: در انتها به‌کمک نتایج به‌دست‌آمده می‌توان اطلاعات مهم را استخراج کرد و متن خلاصه‌ای با اندازه‌ای کوچک‌تر به‌دست آورد.

الگوریتم‌های خلاصه‌سازی متنی کاربردهای فراوانی دارند و هرروزه به کاربردهای آن‌ها اضافه می‌شود. برای مثال، چند نمونه از کاربردهای آن شامل موارد زیر است[۱۲]:

- **استخراج و خلاصه‌سازی داده‌ای از اخبار**: همه‌روزه اتفاق‌های متفاوتی در دنیا در حال رخ‌دادن و بشر برای یافتن اطلاعات خاص از این حجم بالای داده‌ای دچار مشکل است. به‌کمک خلاصه‌سازی این مشکل را می‌توان رفع کرد[۱۳].
- **خلاصه‌سازی فصل‌های کتب مرجع**: کتاب‌های مرجع حاوی عناوینی طولانی هستند که با جزییات توضیح داده شده‌اند. به‌کمک خلاصه‌سازی چندسنده می‌توان خلاصه‌ای دقیق و کوتاه از این داده‌ها ایجاد کرد.
- **خلاصه‌سازی سخنرانی**: این نظریه را می‌توان در خلاصه‌سازی فرایند تماس‌ها، گزارش‌های خبری، گفت‌وگوها و... استفاده کرد[۱۴].

با رشد منابع دیجیتال، توصیهٔ برچسب توجه زیادی را به خود جلب کرده‌است. هدف یک سامانهٔ توصیهٔ برچسب، ارائهٔ مجموعه‌ای از برچسب‌ها برای یک قطعه متن است تا فرایند برچسب‌گذاری را که به‌صورت دستی توسط کاربر انجام می‌شود، آسان کند[۱۵]. این برچسب‌ها قابلیت‌های موتورهای جست‌وجو را برای پیمایش، سازمان‌دهی و جست‌وجوی محتوا را افزایش می‌دهند؛ بااین‌حال، برچسب‌گذاری دستی متن زمان‌بر و پر زحمت است.

سامانهٔ توصیه، یک رویکرد کارآمد برای غلبه بر مشکل اضافه‌بار اطلاعات است. توصیهٔ برچسب به کاربران اجازه می‌دهد تا صفحات وب، موسیقی و مقالات را با واژه‌های کلیدی حاشیه‌نویسی کنند و برای جست‌وجوی محتوای چندرسانه‌ای مفید است.

کاربران می‌توانند از هر واژگانی برای برچسب‌گذاری موارد مورد علاقهٔ خود استفاده کنند؛ بنابراین، در مقایسه با ماتریس امتیازدهی در سامانه‌های توصیهٔ عمومی، اطلاعات برچسب می‌تواند علایق و عادات کاربران را با دقت بیشتری بیان کند؛ علاوه‌براین، برای هر موردی، کاربران مختلف اغلب برچسب‌های متفاوتی ارائه می‌دهند، که می‌تواند محتوای نمونه را از چند نما توصیف و به کاربران در جست‌وجوی دقیق و ارائهٔ توصیهٔ باکیفیت کمک کند[۱۶].

در ابتدای راه توصیهٔ برچسب سعی شد تا از روش‌های سنتی الگوریتم‌های توصیه در برچسب‌گذاری استفاده شود. یکی از معروف‌ترین روش‌ها روش پالایهٔ مشارکتی مبتنی بر برچسب[4] است[۱۷]؛ کاربران می‌توانند با هر واژهٔ دل‌خواه و تصادفی و به هر تعداد برچسب‌گذاری داشته باشند که این منجر به مشکل افزونگی و ابهام می‌شود. برای رفع این مشکلات، انواع دیگری از تکنیک‌ها مانند رویکردهای مبتنی بر خوشه‌بندی [۱۸]، رویکردهای مبتنی بر ماتریس [۱۹] و رویکردهای مبتنی بر نمودار [۲۰] معرفی شده‌اند.

در این مقاله سعی شده‌است به‌کمک الگوریتم K-NN و ترکیب آن با مدل BoW و سایر مدل‌های معنایی و به‌کمک خلاصه‌سازی متن، برچسب‌های مناسب جهت استفاده و ارائهٔ پیشنهاد در شبکه‌های اجتماعی استخراج شود.

اهداف اصلی این مقاله:

- انتخاب بهترین برچسب جهت ارائهٔ توصیه در شبکه‌های اجتماعی انجام گیرد.
- استفاده از مدل‌های معنایی کاوش متن و خلاصه‌سازی متن، برای به‌دست‌آوردن بهترین برچسب در شبکه‌های اجتماعی انجام گیرد.

[1] Part-of-speech
[2] Natural Language Processing (NLP)
[3] Python Natural Language Toolkit
[4] Tag Based Collaborative Filtering

[DOI: 10.61186/jsdp.21.4.15]

- نتیجه بر روی دو مجموعه‌دادگان عمومی TPA [۲۱]، AG[۲۲] ارزیابی شود تا بر اساس آن کارایی راه‌کار پیشنهادی مورد بررسی قرار گیرد.

مقاله به‌شرح زیر سازماندهی شده‌است: بخش دو پیشینه و کارهای مرتبط را بررسی، بخش سه چهارچوب پیشنهادی را تشریح می‌کند و بخش چهار به بررسی ارزیابی روش پیشنهادی می‌پردازد؛ درنهایت، نتیجه‌گیری در بخش پنج ارائه شده‌است.

۲- مرور ادبیات موضوع

در ابتدا برای مروری کلی در روش‌های به‌کاررفته در توصیهٔ برچسب به معرفی الگوریتم‌های خلاصه‌سازی متن پرداخته و در ادامه پژوهش‌هایی که بر روی توصیهٔ برچسب و بهبود آن کار کرده‌اند؛ معرفی می‌شوند.

۲-۱-خلاصه‌سازی متن

خلاصه‌سازی متن به‌کمک فرایند پردازش زبان طبیعی، سعی در تبدیل متن به یک نسخهٔ کوتاه‌تر دارد. مطالعات زیادی، چالش‌های خلاصه‌سازی خودکار متن را در نیم‌قرن اخیر بررسی کرده‌اند[۲۳]. مقالاتی که به‌صورت پیش‌گام در این زمینه کار کرده‌اند و در ادامه توانستند از این خلاصه‌سازی در صفحات وب بهره ببرند، به‌ترتیب بیان خواهند شد.

میلاد مرادی و همکاران [۲۴] از روش خلاصه‌سازی بیزی[1] برای اسناد متنی پزشکی استفاده کردند. خلاصه‌سازی بیزی، در ابتدا متن ورودی را به مفاهیم سامانهٔ زبانی یک‌پارچهٔ پزشکی (UMLS[2]) نگاشت و سپس مهم‌ترین آن‌ها را برای طبقه‌بندی ویژگی‌ها استفاده می‌کند. از شش نظریهٔ متفاوت انتخاب ویژگی برای شناسایی مفاهیم مهم استفاده شد و بر اساس توزیع این مفاهیم، موارد با بار اطلاعاتی بالاتر انتخاب شدند. به‌کمک خلاصه‌سازی بیزی، کارایی خلاصه‌ساز بهبود یافت. با توجه به اینکه روش بیزی روشی آماری و از تخمین استفاده می‌کند توانسته است در زمینه پزشکی موفق باشد، اما نمی‌توان در کل و در روش‌های عمومی‌تر از آن بهره برد.

روتاری[3] و بالبانتاری[4] [۲۵] روش جدیدی با نام جست‌وجوی فاخته را معرفی کردند که بر اساس خلاصه‌سازی چندسنده است. هدف آن‌ها نمایش مشکلات و چالش‌های موجود در این روش بوده‌است. جست‌وجوی فاخته یا به‌اختصار CS [5] یکی از الگوریتم‌های فراابتکاری است که از پرنده‌ای با نام فاخته الهام گرفته‌است. فاخته‌های بالغ، تخم‌های خود را در لانهٔ سایر پرندگان یا حیوانات قرار می‌دهند. لانه‌ای که یکی از تخم‌ها را در خود داشته باشد، درواقع یک راه حل و هر فاخته تنها می‌تواند یک تخم را که جدیدترین و قوی‌ترین پاسخ است در یک لانه قرار دهد. الگوریتم CS می‌تواند از ساده‌ترین قالب، یعنی زمانی‌که هر لانه تنها یک تخم دارد تا جایی که هر لانه بتواند چندین تخم داشته باشد (که نشان‌دهندهٔ مجموعه‌ای از پاسخ‌هاست) پیشروی داشته باشد. پس از پیش‌پردازش، امتیاز ارزش معنایی جمله محاسبه و سپس جست‌وجوی فاخته پیاده‌سازی و سرانجام بهینه‌ترین جملات انتخاب شدند تا متن خلاصهٔ نهایی را شکل دهند. این روش تنها درخصوص داده‌های عمومی به خوبی عمل می‌کند و در زمانی که داده‌ها برای شرایط موضوعی و سفارشی متمرکز شوند (ورزشی، درمانی و...) خروجی خوبی نخواهد داشت.

روتاری و بالبانتاری [۲۶] در مقاله‌ای دیگر، با استفاده از الگوریتم بهینه‌ساز ازدحام ذرات[6] یک خلاصه‌سازی عمومی را برای اسناد تک متن معرفی کردند. این الگوریتم پوشش محتوایی و ویژگی افزونگی را به‌عنوان جنبه‌های مهم خلاصه‌سازی در نظر می‌گیرد؛ درنهایت نتایج رضایت‌بخش بوده‌است. دقت این روش بالاست، اما زمانی‌که حجم داده زیاد و نوفه بیشتر شود، دقت به میزان قابل توجهی افت خواهد کرد.

در ادامه تعدادی روش که از یادگیری عمیق و روش‌های خوشه‌بندی در خلاصه‌سازی متن استفاده کرده‌اند معرفی می‌شوند.

آزادانی[7] و همکاران [۲۷] از سامانهٔ یک‌پارچه زبان پزشکی برای ایجاد مدل مفهومی اسناد و نگارش آن‌ها به مفاهیم استفاده کردند. این روش موارد پرتکرار را کشف و هم‌بستگی میان نظریه‌های مختلف را پیدا کرده‌است. از همبستگی‌ها برای پیشنهاد تابع تشابه استفاده می‌کند تا یک گراف ایجاد شود؛ سپس، خلاصه‌ساز الگوریتم خوشه‌بندی را بر اساس درخت پوشای کمینه به‌کار می‌برد تا زیرمجموعه‌های مختلف اسناد را پیدا کند. سرانجام خلاصهٔ نهایی را به‌کمک انتخاب جملاتی با بار اطلاعاتی بالا و مرتبط‌ترین جملات، از میان زیرمجموعه‌ها در متن ایجاد می‌کند. ارزیابی خودکار بر روی تعداد بالایی از داده‌ها با استفاده از مقیاس‌های روژ[8] انجام شد. نتایج نشان داد که سامانهٔ پیشنهادی کارایی بالاتری در برابر نظریه‌های مرسوم داشته‌است.

[1] Bayesian
[2] Unified Medical Language System
[3] Rautray
[4] Balabantaray

[5] Cuckoo search
[6] Particle Swarm Optimization Algorithm
[7] Azadani
[8] Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

[DOI: 10.61186/jsdp.21.4.15]

بارالیس[1] و همکاران [۲۸] از کاوش مکرر داده‌ها و روش معروف یافتن الگوهای تکراری در منبع داده استفاده کردند. در این روش با جست‌وجو در میان کل سند، داده‌هایی که به‌صورت پرتکرار در متن آورده شده بودند یا الگوی مفهومی تکراری داشتند، شناسایی و به‌عنوان خروجی خلاصه استخراج شدند. نتایج، برتری روش پیشنهادی را نشان داده است. با وجود این با افزایش حجم داده‌های ورودی سرعت الگوریتم به‌شدت کاهش می‌یابد.

تهالینو[2] و آمانچیو[3] [۲۹] کارایی روش چندلایه‌ای را برای انتخاب جملات مرتبط در خلاصه‌سازی چندسنده ارزیابی کردند. در مدل به‌کار رفته، گره‌ها نشان‌دهندهٔ جملاتند و یال‌ها بر اساس تعداد واژه‌های اشتراکی بین جملات ایجاد شدند. بر خلاف مطالعات پیشین در خلاصه‌سازی چندسنده، در این روش تمایزی بین یال‌های اتصالی از یک سند و آن‌هایی که اسناد مختلف را وصل می‌کردند وجود داشت. نتایج نشان داد که این تمایز بین لایه‌های داخلی و خارجی در نمایش چند لایه‌ای می‌تواند کیفیت خلاصه‌های ایجاد شده را بهبود بخشد.

افشاری‌زاده و همکاران [۳۰] خلاصه‌سازی متنی را بر اساس روش جست‌وجوگرا (کاربرگرا) با استخراج مهم‌ترین جملات ارائه دادند؛ به همین منظور، چندین ویژگی از جملات استخراج شدند که هر کدام اهمیت جملات را از جنبه‌ای خاص سنجیدند. در این مقاله یازده مورد از بهترین ویژگی‌ها از هر جمله استخراج شدند. این مقاله نشان داد اگر ویژگی‌های مناسب‌تری استفاده شوند، منجر به خلاصه‌های بهتری خواهند شد.

۲-۲- توصیه‌کنندهٔ برچسب

روش‌های پیشنهادی توصیهٔ برچسب را به روش پالایهٔ مشارکتی[4] و روش مبتنی بر محتوا تقسیم می‌کنیم[۳۱]. ایدهٔ کلیدی روش پالایهٔ مشارکتی استفاده از اطلاعات رتبه‌بندی تاریخی است. فنگ[5] و وانگ[6] یک سامانهٔ برچسب‌گذاری اجتماعی را به‌صورت نموداری مدل کرده و با یادگیری وزن گره‌ها و یال‌ها در نمودار، برچسب‌ها را توصیه کردند[۳۲]. فانگ[7] و همکاران روشی جدید برای توصیهٔ برچسب شخصی پیشنهاد کردند. این روش یک توسعهٔ غیرخطی است[۳۳]. ژائو[8] و همکاران روابط در برچسب‌گذاری داده‌ها را به‌عنوان یک نمودار ناهمگن مدل کرده و یک چهارچوب الگوریتمی رتبه‌بندی برای توصیهٔ برچسب پیشنهاد کردند[۳۴]. بر خلاف روش‌های پالایهٔ مشارکتی، روش مبتنی بر محتوا، داده را به‌عنوان ورودی می‌گیرد؛ بنابراین می‌تواند در توصیهٔ برچسب برای محتوای جدید استفاده شود و از مشکلات روش‌های پالایهٔ مشارکتی اجتناب کند. یکی از تکنیک‌های برچسب‌گذاری مبتنی بر محتوا، استفاده از مدل‌های تولیدی است. کرستل[9] و همکاران نوعی مدل تخصیص دیریکله پنهان[10] را برای استخراج یک ساختار موضوعی مشترک از برچسب‌گذاری مشترک چندین کاربر معرفی کردند که توانست با استفاده از مفاهیم نتایج خوبی را نشان دهد[۳۵].

از سوی دیگر، برخی تحقیقات وجود دارد که توصیه برچسب را از طریق مدل مبتنی بر شبکه عصبی عمیق و سپس طبقه‌بندی نمایش به برچسب‌های مختلف، محقق می‌کنند. وستون[11] و همکاران یک معماری عمیق مبتنی بر شبکه‌های عصبی کانولوشنال یا CNN[12] پیشنهاد کردند [۳۶]. این مدل واژه‌ها و همچنین کل پست‌های متنی را در لایه‌های میانی معماری عمیق CNN نشان داده‌است و خروجی معنای مناسبی داشته‌است.

کارهای اخیر با ادغام روش‌های یادگیری عمیق برای رسیدگی به مشکلات بالا استفاده می‌کنند که عملکرد چشم‌گیری را در نتیجه نهایی نشان داده‌است.

آیمن[13] و همکاران از سامانه‌های توصیه‌کنندهٔ پالایهٔ مشترک[14] برای ارائهٔ توصیه‌های مربوط به برچسب ارائه کردند. در روش پیشنهادی بر اساس برچسب‌های معنایی، تشابه میان کاربران را به‌کمک فاصله‌های معنایی ایجادشده در برچسب‌های موجود در پست‌های افراد پیدا می‌کند. این روش توانسته‌است شباهت معنایی میان کاربران را بر اساس نظریهٔ رنگ‌ها بهتر نشان دهد. نتایج آزمایش‌ها بر روی این روش برتری آن را نسبت به سایر روش‌های پالایهٔ مشترک نشان داده‌است[۳۷].

[1] Baralis
[2] Tohalino
[3] Amancio
[4] Collaborative Filtering
[5] Wei Feng
[6] Jianyong Wang
[7] Xiaomin Fang
[8] Wei Zhao
[9] Ralf Krestel
[10] Latent Dirichlet Allocation model
[11] Jason Weston
[12] Convolutional neural networks
[13] Ayman
[14] Collaborative Filtering Recommender Systems

[DOI: 10.61186/jsdp.21.4.15]

ژانگ[1] و همکاران یک روش پیشنهاددهندهٔ برچسب برای متون ارائه دادند[۱]. در این مقاله یک روش ارائه معرفی شد که با ترکیب ارائهٔ معنایی و مدل عنوان به یک راه‌کار موفق برای توصیهٔ برچسب دست یافته‌است. با بررسی و مقایسهٔ این روش با سایر روش‌های مشابه، برتری این روش مشخص شد.

حمیدزاده و مرادی روش جدیدی برای ارائهٔ توصیه ارائه دادند[۳۸]. از الگوریتم خوشه‌بندی فازی C-میانگین مرتب‌شده و الگوریتم تکاملی ازدحام ذرات تطبیقی آشوبی برای خوشه‌بندی کاربران استفاده شده‌است. هدف روش پیشنهادی بهبود میزان خطای پیش‌بینی در مجموعه‌داده‌های حجیم با پراکندگی زیاد و کاهش تأثیر داده‌های پرت و نوفه است. برای ارزیابی و اثبات کارایی روش پیشنهادی، آزمایش‌هایی روی پایگاه داده‌های واقعی اجرا شد. نتایج آزمایش‌ها نشان‌دهندهٔ برتری روش پیشنهادی نسبت به روش‌های مرز دانش بر اساس معیارهای میانگین خطای مطلق، جذر میانگین مربعات خطا، نرخ صحت و زمان محاسباتی است.

بحرانی و دیگران روش جدید توصیه‌دهنده ارائه دادند[۳۹]. در سامانهٔ پیشنهادگر ترکیبی پیشنهادی، از یک سامانهٔ دو مرحله‌ای استفاده کردند که در مرحلهٔ نخست، دو مدل پیش‌بینی‌های خود را انجام داده، سپس در مرحلهٔ دوم به‌وسیلهٔ یک مؤلفهٔ ترکیب‌گر، نتایج دو بخش مرحلهٔ اول را با یکدیگر ترکیب کرده‌است تا خروجی نهایی را به‌دست آورد. نتایج مقایسهٔ روش پیشنهادی با برخی روش‌های مشابه حاکی از آن است که این روش به نسبت سایر روش‌ها، کندتر، اما از آن‌ها دقیق‌تر است.

پن[2] و همکاران به‌کمک روش K-NN سعی در بهبود مقالهٔ پیشین خود داشته و توانستند الگوریتم توصیهٔ برچسب با صحت بالا ارائه دهند[۴۰]. در این روش به‌کمک الگوریتم خوشه‌بندی K-NN در ابتدا برچسب‌ها رتبه‌بندی شدند و سپس برچسب‌هایی با رتبهٔ بالاتر به‌عنوان خروجی انتخاب شدند با اینکه این روش نتیجهٔ خوبی داشت، اما به‌دلیل بهینه‌نبودن داده‌های ورودی نمی‌تواند در مواجه با حجم بالای داده‌ها به خوبی عمل کند.

جیمل[3] و همکاران نیز از الگوریتم K-NN برای توصیهٔ برچسب استفاده کردند[۴۱]. در ابتدا از دو تکنیک کاهش داده‌ها، پردازش هستهٔ-p و کاهش هبی[4] بهره بردند سپس به‌کمک الگوریتم وفقی K-NN توانستند برچسب‌های مناسبی پیشنهاد دهند و صحت بالایی داشته‌است، اما از این روش بر روی استخراج ویژگی و بهبود آن برای بالابردن صحت نتیجهٔ نهایی استفاده نکرده‌است. موردی که سعی شد در الگوریتم پیشنهادی مورد توجه قرار گیرد.

با توجه به تمام موارد قید شده این مقاله سعی دارد تا به‌کمک‌گرفتن از روش‌های خلاصه‌سازی متن و بهبود داده‌های ورودی، یک روش توصیهٔ برچسب با روش خوشه‌بندی ارائه دهد که صحت و دقت روش‌های پیشین را افزایش داده و بتواند در پیشنهاد برچسب در پایگاه داده‌های عظیم موفق باشد.

۳- راه‌کار پیشنهادی

در شکل (۱) مدلی کلی از راه‌کار پیشنهادی قابل مشاهده است. در ابتدا به‌کمک مدل BoW تجزیهٔ واژه‌ها و ساده‌سازی و استخراج ریشهٔ واژه‌ها به‌دست می‌آید. در ادامه با ترکیب الگوریتم ESA و K-NN بهترین برچسب‌ها، به‌عنوان برچسب کاندید انتخاب می‌شوند. در ادامه هر یک از گام‌های راه‌کار پیشنهادی به‌صورت جزئی توضیح داده شده‌است.

۳-۱- K-نزدیک‌ترین همسایه

K-NN یکی از روش‌های معروف دسته‌بندی است. در این روش تصمیم‌گیری این‌که یک نمونهٔ جدید در کدام دسته قرار گیرد با بررسی تعدادی (k) از شبیه‌ترین نمونه‌ها یا همسایه‌ها انجام می‌شود. در بین این k نمونه، تعداد نمونه‌ها برای هر دسته شمرده می‌شود و نمونهٔ جدید به دسته‌ای که تعداد بیشتری از همسایه‌ها به آن تعلق دارند نسبت داده می‌شود.

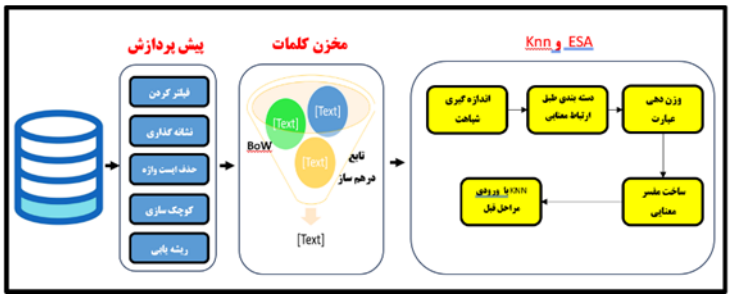
اولین کار برای استفاده از K-NN یافتن معیاری برای شباهت یا فاصلهٔ بین صفات در داده‌ها و محاسبهٔ آن است. در حالی که این عمل برای داده‌های عددی آسان است، متغیرهای دسته‌ای نیاز به برخورد خاصی دارند. هنگامی که فاصلهٔ بین نمونه‌های مختلف را توانستیم اندازه بگیریم، می‌توانیم مجموعه نمونه‌هایی که پیش‌تر دسته‌بندی شده اند را به‌عنوان پایهٔ دسته‌بندی نمونه‌های جدید استفاده کنیم. در روش کلاسیک، فاصلهٔ بین دو نمونهٔ $X=(x_1,\ldots,x_n)$ و $Y=(y_1,\ldots,y_n)$ با استفاده از فاصلهٔ اقلیدسی از طریق فرمول (۱) محاسبه می‌شود:

$$d(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^n} \qquad (۱)$$

فهم مدل‌های K-NN هنگامی که تعداد متغیرهای پیش‌بینی‌کننده کم است، بسیار ساده است. این الگوریتم همچنین برای ساخت مدل‌هایی مانند متن که شامل انواع دادهٔ غیراستاندارد هستند، بسیار مفید است. تنها نیاز برای انواع دادهٔ جدید وجود یک معیار مناسب شباهت است.

[1] Shangru Zhong
[2] Rong Pan
[3] Jonathan Gemmell
[4] Hebbian deflation

[DOI: 10.61186/jsdp.21.4.15]

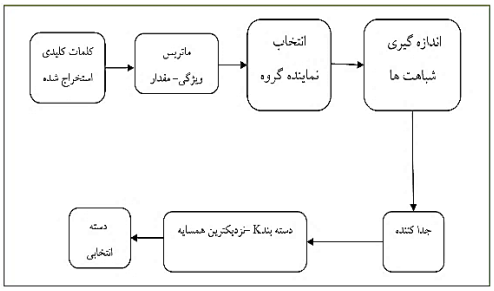


(شکل-۱): مدلی کلی از راه‌کار پیشنهادی
(Figure-1): A general model of the proposed solution

قالب مفهومی الگوریتم ترکیبی K-NN در شکل (۲) قابل مشاهده است. یک مجموعه از سندهای متنی در سامانه به‌کار برده شده‌است. برچسب‌ها برای نشان‌دادن دسته‌ای که سند متنی به آن بستگی دارد، استفاده می‌شود. در این پیاده‌سازی تمام سندهای متعلق به مجموعه‌داده باید برای یادگیری سامانه و سپس آزمایش آن، برچسب‌گذاری شود.

۳-۱-۱-پیش‌پردازش

پیش‌پردازش، عملیات آماده‌سازی متن خام برای فرایندهای کاوش متنی است. این عمل برای کاهش نوفه در داده مناسب است. هدف اصلی آن تبدیل دادهٔ اصلی به قالبی است که برای سامانهٔ ماشینی قابل درک باشد[۴۲]. فرایند پیش‌پردازش شامل نشان‌گذاری، پالایه، ریشه‌یابی، حذف ایست‌واژه‌ها[1] و... است که در این مقاله، بعضی از آن‌ها استفاده شده‌اند.

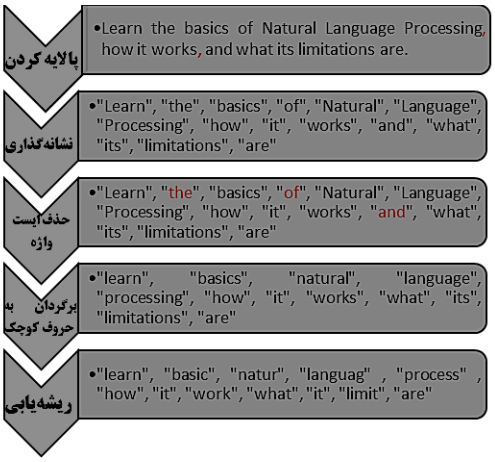


(شکل-۲): مدل K-NN
(Figure-2): K-NN model

در شکل (۳) نمونهٔ پیش‌پردازش برای جملهٔ **"Learn the basics of Natural Language Processing, how it works, and what its limitations are"** نشان داده شده‌است.

پالایه‌کردن: فرایند حذف اطلاعات و واژه‌های کم‌اهمیت‌تر برای کاهش محاسبات است. تمام علامت‌ها، نویسه‌های خاص و... از داده خارج شده‌است.

نشانه‌گذاری: فرایند تقسیم‌بندی جملات طولانی به قسمت‌های کوچک‌تر، به‌صورت واژه‌به‌واژه است؛ یعنی جملات به مجموعه‌ای از نشانه‌های جدا که به‌وسیلهٔ فاصله‌ها از هم جدا می‌شوند، تقسیم شده و می‌توانند برای پردازش و درک بیشتر استفاده شوند. نشانه‌ها می‌توانند تک‌واژه، واژه‌های کلیدی، عبارات، شناسه‌ها و غیره باشند. در این فرایند، نشانه‌ها یا واژه‌های با فضای خالی، خطوط شکسته یا علائم نگارشی از هم جدا می‌شوند.



(شکل-3): مدل پیش‌پردازش
(Figure-2): Pre-processing model

حذف ایست‌واژه‌ها: اگرچه فرکانس واژه‌ها اهمیت آن‌ها را نشان می‌دهد، اما واژه‌های اضافی، اتصال‌دهنده‌ها و... که در تمام اسناد دیده می‌شوند، بالاترین فرکانس را دارند، اما حاوی اطلاعات مهمی نیستند. همهٔ آن‌ها درواقع ایست‌واژه به‌حساب می‌آیند و باید از داده حذف شوند. در این بخش با تشخیص و تعیین یک‌سری ایست‌واژه، همگی آن‌ها از مجموعه‌داده حذف می‌شوند. این ایست‌واژه‌ها اغلب حروف اضافه‌اند؛ مانند "a"، "of" و...

برگردان به حروف کوچک: برای جلوگیری از سربار اضافه و تحلیل برتر، تمامی واژه‌ها به حروف کوچک تبدیل می‌شوند تا از سربار اضافی در محاسبات جلوگیری شود.

ریشه‌یابی: در این فرایند تمام واژه‌ها به ریشه‌های خود خلاصه می‌شوند. با این کار سربار محاسباتی

[1] Stop-Word

[DOI: 10.61186/jsdp.21.4.15]

کم‌تر و دقت نتیجه را بالاتر می‌برد؛ چرا که واژه‌هایی که یک معنا دارند، اما با شکل‌های مختلف در جملات قرار گرفتند، به ریشهٔ خود تبدیل می‌شوند و هم ارزش واژهٔ خود را بالا می‌برند (تکرار بیشتری دارند) و هم تعداد محاسبات را در ادامه کاهش می‌دهند.

در پایان این مرحله داده‌های ورودی پالایش، لغات زائد، واژه‌های اضافه و موارد بی‌اهمیت حذف می‌شوند و واژه‌های با معنای بیشتر باقی می‌مانند تا در مرحلهٔ بعدی کتابخانهٔ واژه‌ها را تشکیل دهند.

(الگوریتم-۱): حذف ایست‌واژه‌ها

(الگوریتم-۱): حذف ایست‌واژه‌ها

(Algorithm-1): StopWord Remover

```
Procedure StopWord Remover
n=size(raw_text);
For i=1:n
   If ismember(raw_text[i] ,StopWords)
     remove raw_text[i]
```

۳-۱-۲- BoW تغییریافته

مخزن واژه‌ها یا BoW، مدلی است که تعداد تکرار واژه‌ها را در اسناد می‌شمارد. در این نظریه، متن به‌صورت بردار $\vec{d}$ از وزن واژه‌ها نشان داده می‌شود یعنی در هر سند $d_i \in D$، d_i برداری با ابعاد بالای وزنی است و هر بعد نشان‌دهندهٔ یک واژهٔ خاص است[۴۳].

ورودی T[i,j] در جدول مربوط به یک مقدار tf-idf از واژهٔ t_i در سند d_j است.

$$T[i,j] = tf(t_i, d_j).\log\frac{n}{df_i} \qquad (۲)$$

که میزان تکرار واژه در آن برابر است با:

$$tf(t_i, d_j) = \begin{cases} 1+\log count(t_i, d_j), & count(t_i, d_j) > 0 \\ 0 & otherwise \end{cases} \qquad (۳)$$

و $df_i = |\{d_k : t_i \in d_k\}$ | برابر است با تعداد اسناد مجموعه که شامل واژهٔ t_i (میزان تکرار سند) است.

مطابق با الگوریتم (۲)، BoW برای تمام متون محاسبه می‌شود. این الگوریتم فرکانس هر واژه (افعال و واژه‌ها) را مشخص و واژه‌های با فرکانس بالاتر را برای ایجاد برچسب جدید، انتخاب می‌کند.

در پایان این مرحله واژه‌ها و تعداد تکرار آن‌ها آماده و به‌عنوان دادهٔ ورودی وارد درهم‌ساز، اما پیش از این کار در مرحلهٔ بعد افعال از سایر واژه‌ها جدا می‌شوند.

(الگوریتم -۲): BoW تغییریافته

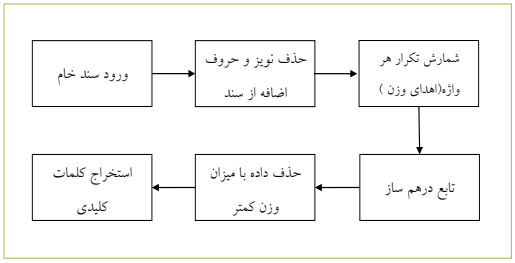
(الگوریتم -۲): BoW تغییریافته

(Algorithm-2): Modified BoW

```
Procedure Modified BoW
 For each role[i]
     [n m]=size(role[i])
     For j=1:n
        For t=1:m
           Calculate the term frequency
         Create a vector of frequency for this word
```

۳-۱-۳- تجزیه‌کنندهٔ فعل

افعال در جملات اهمیت بالایی دارند، اما گاهی فرکانس آن‌ها چنان پایین است که حذف می‌شوند و گاهی چنان بالا که واژه‌های دیگر با فرکانس پایین را حذف می‌کنند. در هر دو صورت، تأثیر معکوسی بر روی نتایج می‌گذارند.



(شکل-۴): مدل BoW
(Figure-4): The BoW model

از سوی دیگر در زبان انگلیسی واژه‌هایی وجود دارند که هم به‌عنوان فعل و هم به‌عنوان اسم کاربرد دارند (مانند Record، Track و...) و استفاده از آن‌ها در جایگاه اشتباه، باعث بالارفتن خطا می‌شود. یک راه حل خوب برای این مشکل، جداسازی افعال از سایر واژه‌هاست؛ به همین منظور فرکانس افعال به‌صورت جداگانه محاسبه می‌شود. در این گام برای هر جمله، فعل‌های آن جدا شده و فرکانسش محاسبه و سپس افعال کلیدی انتخاب شده و برای حذف اثر منفی، از داده حذف می‌شوند. نمای کلی از منطق ارائه شده در شکل (۴)، قابل مشاهده است.

حال واژه‌ها و افعال کلیدی آماده شده و در یک کتابخانه قرار می‌گیرند و در مرحلهٔ بعد به‌کمک تابع درهم‌سازی برای ورودی تابع KNN آماده شوند.

۳-۱-۴- تابع درهم‌سازی

یک روش رایج برای استفاده از لغت‌نامه‌ها بهره‌گیری از تابع درهم‌سازی است که در آن واژه‌ها مستقیم به نشان‌هایی نگاشت می‌شوند[۴۴]. با نگاشت واژه‌ها به نشانه‌ها از طریق یک تابع درهم‌ساز نیاز به هیچ حافظه‌ای برای دیکشنری وجود نخواهد داشت. تصادم درهم‌سازها به‌طور معمول از طریق باکت‌های درهم‌ساز باعث آزادسازی حافظه می‌شوند و درهم‌سازی به‌علاوه می‌تواند موجب ساده‌سازی مدل BOW و بهبود مقیاس‌پذیری در آن‌ها شود.

[DOI: 10.61186/jsdp.21.4.15]

(الگوریتم-۳): تجزیه‌کنندهٔ فعل

(Algorithm-4): Verb parser

```
Procedure Verb parser
Verbs=chunk(raw_text,verb,XXX.format)
m=size(role)
n =size(verbs)
For i=1: m
  For j=1:n
     If ismember(verbs[j], role[i])
        Verbfrequency[i]=ModifiedBow(verbs[j])
         raw-text=remove(raw_text,verbs[j])
```

۳-۱-۵- انتخاب نمایندهٔ گروه

در این الگوریتم هر دادهٔ جدیدی که وارد می‌شود با تمامی عناصر موجود در تمام گروه‌ها مقایسه می‌شود و بر اساس نزدیکی فاصله‌اش با آن‌ها در یک گروه قرار می‌گیرد. مسلم است این روش باعث پایین‌آمدن کارایی و کندشدن سامانه می‌شود. برای رفع این مشکل در راه‌کار پیشنهادی در هر گروه یک عنصر به‌عنوان نمایندهٔ[1] گروه انتخاب می‌شود. این نماینده، داده‌ای است که نسبت به سایر عناصر مرکزیت بیشتری و نسبت به تمام عناصر گروه فاصلهٔ کمتری داشته باشد و در مرحله بعد با همسایهٔ خود از لحاظ شباهت اندازه‌گیری شود. حال هر عنصر که وارد پایگاه داده می‌شود در هنگام قرارگرفتن در گروه‌ها با نمایندهٔ هر گروه سنجیده می‌شود و در گروهی قرار می‌گیرد که به نمایندهٔ آن گروه نزدیک‌تر است. به جهت به‌روزبودن یک گروه، بعد از هر چند بار اضافه‌شدن یک عنصر جدید به گروه دوباره میان عناصر مقایسه‌ای انجام‌گرفته و مرکز گروه جدیدی برگزیده می‌شود.

۳-۱-۶- اندازه‌گیری شباهت‌ها

رایج‌ترین روش برای اندازه‌گیری شباهت بین سندهای نمایش داده‌شده به‌وسیلهٔ بردارها در فضای مختصات r بعدی که r تعداد عبارت‌ها در فضای ویژگی است، اندازه‌گیری شباهت کسینوسی است. اگر A و B به‌عنوان بردارهای نمایش‌دهندهٔ سندهای j و k فرض شوند، می‌توان شباهت بین A و B را با استفاده از فرمول محاسبه کرد. باید توجه داشت A و B ممکن است، دارای طول‌های متفاوتی باشند.

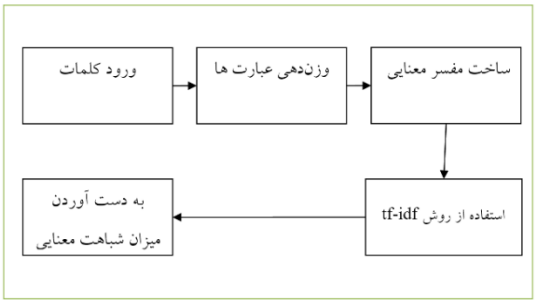
$$sim(A,B)=\frac{\sum_{i=1}^{r} w_{ij}\times w_{ik}}{\sqrt{\sum_{i=1}^{r} w_{ij}^{\,2}}\times\sqrt{\sum_{i=1}^{r} w_{ik}^{\,2}}} \qquad (۴)$$

در اینجا برای اندازه‌گیری شباهت بین واژه‌ها از الگوریتم معنایی ESA کمک گرفته شده‌است.

[1] Medoid

۳-۱-۷- دسته‌بندی بر اساس ارتباط معنایی

در این گام ابتدا سعی می‌شود تا ارتباط معنایی بین واژه‌ها پیدا شود. به‌کمک این ارتباط معنایی می‌توان هم واژه‌های مهم و اصلی را به‌راحتی استخراج کرد و هم به نوع محتوای متن پی برد؛ درنتیجهٔ این عملیات، یک‌سری واژه‌های مرتبط به‌دست می‌آید که همگی در یک گروه قرار می‌گیرند. گروه‌ها تشکیل می‌شوند تا در ادامه جست‌وجو بر اساس آن‌ها انجام شود. روندنمای الگوریتم ESA در شکل (۵) قابل مشاهده‌است.



(شکل-۵): روندنمای الگوریتم ESA
(Figure-5): The ESA flowchart

۳-۱-۸- وزن‌دهی عبارت‌ها

اگر چه روش‌های وزن‌دهی متفاوتی برای اندیس‌گذاری وجود دارد، اما همهٔ آن‌ها در دو مورد زیر مشترک‌اند:

- هر چه تعداد دفعات رخداد یک عبارت در یک سند که متعلق به یک دسته است، بیشتر باشد، آن عبارت ارتباط بیشتری با دستهٔ مذکور دارد.
- هر چه عبارت در سندهای مختلفی که نشان‌دهندهٔ دسته‌های متفاوتی هستند، بیشتر تکرار شود، کمتر برای تمایز بین دسته‌های مختلف موجود، مناسب است.

برای اینکه کلیهٔ اسناد دارای یک مجموعه ویژگی باشند، ویژگی‌هایی را که در بعضی اسناد موجود نبود، برای آن دسته از اسناد با وزن صفر در نظر گرفته شده‌است. در مرحلهٔ آزمایش کلیهٔ اسناد دارای مجموعه ویژگی‌های یکسان فرض شدند و سنجش فاصلهٔ بین ویژگی‌ها برای ویژگی‌های موجود در مجموعهٔ مذکور انجام شد.

۳-۱-۹- ساخت مفسر معنایی

درصورتی‌که مفاهیم $C_1,\ldots\ldots,C_n$ و مجموعه‌ای از اسناد $d_1 , \ldots\ldots , d_n$ به صورت ورودی ارائه شوند از روی آن‌ها جدول پراکندگی T ایجاد می‌شود که در آن هر یک از n ستون مربوط به یک مفهوم خواهند بود و هر یک از سطرها مرتبط با یک واژه‌اند که در $\bigcup_{i=1\ldots.n} d_i$ رخ می‌دهد. همان‌طورکه گفته شد، ورودی T[i,j] در جدول مربوط به یک مقدار tf-idf از واژهٔ t_i در سند d_j است. سرانجام نرمال‌سازی کسینوسی برای هر سطر اعمال می‌شود تا تفاوت میان طول اسناد از بین رود:

[DOI: 10.61186/jsdp.21.4.15]

$$T[i,j] \longleftarrow \frac{T[i,j]}{\sqrt{\sum_{l=1}^{r} T[i,j]^2}} \qquad (۵)$$

(جدول-۱) : مقایسهٔ کارایی روش پیشنهادی با روش‌های مختلف

(Table-1): Comparing the efficiency of the proposed method with different methods

روش	TPA			AG		
	Macro-F1	Macro-R	Macro-P	Macro-F1	Macro-R	Macro-P
AdaBoost[49]	۰/۷۵۱	۰/۷۲۱	۰/۷۳۱	۰/۷۹۹	۰/۷۸۰	۰/۷۹۹
RF[50]	۰/۷۴۴	۰/۷۲۵	۰/۷۳۲	۰/۷۶۹	۰/۷۶۹	۰/۷۶۸
GBDT[51]	۰/۸۱۱	۰/۷۸۹	۰/۷۹۷	۰/۸۲۰	۰/۸۲۱	۰/۸۲۰
LSTM[52]	۰/۸۰۵	۰/۷۹۷	۰/۷۹۸	۰/۸۶۱	۰/۸۶۲	۰/۸۶۱
BiLSTM[53][۱]	۰/۸۱۵	۰/۸۱۱	۰/۸۱۰	۰/۸۸۲	۰/۸۸۰	۰/۸۸۱
Att-BiLSTM[54]	۰/۸۱۹	۰/۸۱۱	۰/۸۱۲	۰/۸۹۱	۰/۸۹۰	۰/۸۹۰
CNN[55]	۰/۸۰۴	۰/۷۹۸	۰/۸۰۰	۰/۹۱۴	۰/۹۰۸	۰/۹۱۱
ABCNN[56]	۰/۸۱۷	۰/۸۱۳	۰/۸۱۱	۰/۹۱۷	۰/۹۱۳	۰/۹۱۴
VD-CNN[57] [۲]	۰/۸۱۳	۰/۸۱۳	۰/۸۰۹	۰/۹۱۳	۰/۹۱۰	۰/۹۱۲
CapsNet[58] [۳]	۰/۸۲۰	۰/۸۱۵	۰/۸۱۴	۰/۹۲۱	۰/۹۱۸	۰/۹۲۰
Capsule-B[59]	۰/۸۱۸	۰/۸۰۶	۰/۸۱۰	۰/۹۲۶	۰/۹۱۹	۰/۹۱۷
CAN[31]	۰/۸۲۹	۰/۸۲۵	۰/۸۲۴	۰/۹۲۶	۰/۹۲۲	۰/۹۲۳
روش پیشنهادی	۰/۸۲۹	۰/۸۳۱	۰/۸۲۹	۰/۹۳۰	۰/۹۳۱	۰/۹۳۰

[۱] Bidirectional LSTM [۲] Very Deep Convolutional Network [۳] Capsule Network

که r تعداد واژگان است. تفسیر معنایی واژهٔ t_i در سطر i جدول T قرار دارد. این یعنی معنی یک واژه از طریق بردار مفاهیم با مقدار tf-idf آن همراه است که نشان‌دهندهٔ میزان مرتبط‌بودن هر مفهوم به یک واژه است. در این گام، برچسب‌های کلیدی انتخاب می‌شوند.

۳-۱-۱۰-مرحلهٔ پایانی

در مرحلهٔ پایانی داده‌های کلیدی و مقادیر به‌دست‌آمده از مراحل قبلی در الگوریتم KNN که در ابتدا توضیح داده شد، قرار گرفته و از میان واژگان کلیدی بهترین موارد انتخاب می‌شوند که می‌توان به‌عنوان توصیهٔ برچسب استفاده شوند.

۴- ارزیابی راه‌کار پیشنهادی

در این بخش به معرفی تنظیمات کلی انجام آزمایش بر روی نتایج حاصل از راه‌کار پیشنهادی پرداخته و مورد بررسی قرار می‌گیرند.

۴-۱- مجموعه‌دادگان

به جهت بررسی و ارزیابی روش پیشنهادی، راه‌کار پیشنهادی بر روی مجموعه‌دادگان عمومی یعنی TPA [21] که توسط Aminer[1] و AG [۲۲] که توسط ComeToMyHead[2] استخراج شده‌اند، مورد بررسی قرار گرفت.

مجموعه‌دادگان TPA شامل ۱۸۴۶ مقالهٔ علمی و پنج نوع مختلف برچسب است. این برچسب‌ها به‌ترتیب از داده‌های پایگاه داده (۵۰۵۹ مورد)، مفاهیم بصری(۴۰۷۴)، مفاهیم نظری(۳۹۹۵)، اطلاعات پزشکی (۳۰۶۶) و مفاهیم کاوش داده (۲۲۷۰) هستند.

مجموعه‌دادگان AG شامل ۱۲۷۶۰۰ مقالهٔ خبری است و چهار نوع برچسب به‌کار رفته‌است. هر برچسب شامل سی‌هزار نمونهٔ آموزشی و ۱۹۰۰ نمونه آزمون است. مفاهیم یا برچسب‌های موجود در این مجموعه‌دادگان همگی به یک میزان یعنی ۳۱۹۰۰ عدد است که در خصوص مفاهیم جهانی، ورزشی، تجاری و علمی هستند. هفتاد درصد داده‌ها در هر مجموعه‌دادگان به‌عنوان دادهٔ آموزشی و سی درصد به‌عنوان دادهٔ آزمون مورد بررسی قرار گرفتند.

۴-۲- ارزیابی کمی

مقیاس F-measure به میزان گسترده‌ای در مقالات مختلف جهت ارزیابی کارایی، مورد استفاده قرار می‌گیرد[۴۵-۴۸] برای محاسبهٔ F-measure در ابتدا مقادیر صحت[3] و پوشش[4] محاسبه می‌شوند. در این حالت صحت و پوشش با فرمول‌های زیر به حساب می‌آید:

$$Precision = \frac{TP}{TP + FP} \qquad (۶)$$

[1] http://resource.aminer.org/lab-datasets/crossdomain/ .
[2] http://www.di.unipi.it/-gulli/AG _ corpus _ of _ news _ articles.html .

[3] Precision
[4] Recall

[DOI: 10.61186/jsdp.21.4.15]

$$\text{Re}call = \frac{TP}{TP + FN} \quad (۷)$$

که TP[1] برابر با مثبت صحیح، FP[2] اشتباه صحیح و FN[3] اشتباه کاذب است. از روی این مقادیر مقدار ماکرو هر دو مطابق فرمول‌های زیر محاسبه شده و از روی آن ماکرو F-measure نیاز به‌دست می‌آید.

$$macro - P = \frac{1}{n}\sum_{i=1}^{n} P_i \quad (۸)$$

$$macro - R = \frac{1}{n}\sum_{i=1}^{n} R_i \quad (۹)$$

بنابراین macro F-measure از فرمول زیر به‌دست می‌آید:

$$macro - F_1 = \frac{2 \times macro - P \times macro - R}{macro - P + macro - R} \times \%100 \quad (۱۰)$$

که در این فرمول‌ها n برابر تعداد کل گروه‌هاست.

۴-۳-جزئیات پیاده‌سازی

راه‌کار پیشنهادی بر روی ماشینی با پردازش‌گر Intel(R) Core(TM) i7-10750H CPU @ 2.60GHz 2.59 GHz و NVIDIA GeForce GTX 1650 Ti پیاده‌سازی شده‌است.

۴-۴- نتایج و تحلیل داده‌ها

روش پیشنهادی با دوازده روش مشابه مورد بررسی قرار گرفت. روش‌های ارائه‌شده درکل در چهار گروه ماشین‌های یادگیری، حافظهٔ بلند-کوتاه مدت (LSTM[4])، شبکهٔ عصبی کانولوشنال (CNN[5]) و مدل کپسول‌محور هستند. با مقایسهٔ روش پیشنهادی مشخص‌شده این روش بالاترین میزان صحت، پوشش و ماکرو f1 را دارد. روش‌های متداول یادگیری ماشین آماری مانند تقویت تطبیقی (AdaBoost[6]) و جنگل تصادفی RF[7] عملکرد ضعیفی داشتند؛ زیرا توانایی نمایش ضعیفی دارند. عملکرد این روش‌ها به‌شدت به ویژگی‌ها وابسته است که کار با آن‌ها زمان‌بر است. روش درخت تصمیم افزایش گرادیان (GBDT[8]) عملکرد بسیار بالاتری نسبت به روش‌های AdaBoost و RF دارد؛ زیرا GBDT درخت‌ها را یکی پس از دیگری می‌سازد، جایی که هر درخت جدید به تصحیح خطاهای ایجادشده به‌وسیلهٔ درخت قبلی آموزش داده‌شده کمک می‌کند.

رویکردهای یادگیری عمیق با اختلاف زیادی از رویکردهای یادگیری ماشینی معمولی (AdaBoost و RF) بهتر عمل کردند؛ چرا که در این روش‌ها نیاز به مدیریت و مهندسی ویژگی‌ها نیست؛ علاوه‌براین، Att-BiLSTM[9] و ABCNN[10] به طور پایدار از روش‌های LSTM و CNN فراتر می‌روند؛ زیرا سازوکار مبتنی بر توجه را اتخاذ می‌کنند. این رویکردهای مبتنی بر توجه، اطلاعات مهمی را از متن ورودی با نظارت بر اطلاعات برچسب دریافت می‌کنند. شبکهٔ کپسول گامی بیشتر در جهت شناسایی وجود ویژگی‌های مهم و رمزگذاری ویژگی‌های آن‌ها در بردارهای کم بعدی بر می‌دارد. این کارایی شبکهٔ کپسول را برای توصیهٔ برچسب تأیید می‌کند و درنهایت روش ACN است که در میان سایر روش‌ها بهتر عمل کرده و پس از روش پیشنهادی بهترین روش است. در روش شبکه‌ای Capsule و Capsule-B سعی می‌شود ویژگی‌ها را شناسایی و آن‌ها را تبدیل به بردارهای دوبعدی کند و درنتیجه از آن برای توصیهٔ برچسب بهینه‌تر بهره برد. ACN همین کار را می‌کند، اما با استفاده از روش‌های ترکیبی توجه‌محور نتیجهٔ نهایی را نسبت به دو روش قبلی بهبود بخشیده‌است. روش پیشنهادی نیز عمل مشابهی دارد، اما به‌جای استفاده از روش‌های توجه‌محور با بهره‌گیری از روش خلاصه‌سازی متن، توانسته ایرادهای موجود در روش‌های پیشین را رفع کرده و بنابراین به نتیجه بهتر و Macro-f1 بالاتری دست یابد.

روش پیشنهادی مبتنی بر خلاصه‌سازی متن است به‌کمک KNN به‌خوبی در برابر نوفه پاسخ داده‌است، اما در کنار آن محتوامحور یا معنامحور نیست و درصورتی‌که متون به‌کاررفته از لغات اختصاری متفاوت و بی‌تکرار استفاده کنند باعث می‌شود در شناسایی لغات کلیدی و در ادامه‌پیداکردن ویژگی‌های مناسب ناتوان باشد که نیاز است در راه‌کار آینده به آن پرداخته شود.

۵- نتیجه‌گیری

با توجه به استفادهٔ بالا از شبکه‌های اجتماعی در دنیای امروز، بیش از هر زمان دیگری به برچسب‌ها برای نظم‌دهی و جست‌وجوی داده‌های عظیم نیاز است، در حال حاضر ایجاد برچسب به‌صورت خودکار به شدت مورد توجه قرار گرفته‌است. توصیهٔ برچسب، از یک منبع متن را می‌توان به‌عنوان مشکل استخراج متن به حساب آورد. در این مقاله سعی شد به‌کمک خلاصه‌سازی متن از روی

[1] True Positive--
[2] False Positive
[3] False Negative
[4] Long short-term memory
[5] Convolutional neural network
[6] Adaptive Boosting
[7] Random Forest
[8] Gradient Boosting Decision Tree

[9] attention-based BiLSTM
[10] attention-based Convolutional Neural Network

[DOI: 10.61186/jsdp.21.4.15]

داده‌ها، یک‌سری واژه‌های کلیدی پیشنهادی ایجاد و به‌کمک آن یک پیشنهاددهندهٔ برچسب ارائه کرد. در روش پیشنهادی به‌کمک مدل مخزن واژه‌های (BoW)، تحلیل معنایی آشکار (ESA) و ترکیب آن‌ها با الگوریتم k-نزدیک‌ترین همسایه (K-NN)، یک‌سری برچسب پیشنهادی برای شبکه‌های اجتماعی ایجاد کرد. راه‌کار پیشنهادی بر روی دو مجموعه‌دادگان عمومی مورد بررسی قرار گرفت و نتایج برتری خود را نسبت به سایر روش‌های مشابه نشان داد.

این راه‌کار بر روی دو مجموعه‌دادگان عمومی انجام شد. در مجموعه‌دادگان TPA به ماکرو-f1، ۸۳٪ و برای مجموعه‌دادگان AG به ماکرو-f1 ۹۳٪ به‌دست آمدند که نسبت به سایر روش‌ها برتری داشته و کارایی بالاتری از خود نشان داده‌است.

۶-مراجع 6-References

[1] Zhong, S., et al., "Topic representation: A novel method of tag recommendation for text, " in *IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, 2017, pp. 671-676.

[2]]Hasan, K. and V. Ng, "Automatic Keyphrase Extraction: A Survey of the State of the Art," *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 1. 2014. pp. 1262-1273.

[3] Messina, A. and M. Montagnuolo., "Fuzzy mining of multimedia genre applied to television archives," *in 2008 IEEE International Conference on Multimedia and Expo*. 2008.

[4] Rahul, S. Rauniyar, and Monika, "A Survey on Deep Learning based Various Methods Analysis of Text Summarization," in *2020 International Conference on Inventive Computation Technologies (ICICT)*, 2020.

[5] Radev, D. and K. McKeown, "Introduction to the Special Issue on Text Summarization," *Computational Linguistics*, vol. 28, pp. 399 - 408, 2002.

[6] Oliveira, H., et al., "Assessing shallow sentence scoring techniques and combinations for single and multi-document summarization," *Expert Systems with Applications*, vol. **65**: p. 68-86, 2016.

[7] Ferreira, R., et al., "Assessing sentence scoring techniques for extractive text summarization," *Expert Systems with Applications*, vol. 40, pp. 5755-5764, 2013.

[8] Kiyoumarsi, F. and F. Esfahani, "Optimizing Persian Text Summarization Based on Fuzzy Logic Approach," *International Conference on Intelligent Building and Management*, pp.264-269 2011.

[9] Fang, C., H. Kesong, and C. Guilin. "An approach to sentence-selection-based text summarization," in *2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. TENCOM '02. Proceedings*, 2002.

[10] Sankarasubramaniam, Y., K. Ramanathan, and S. Ghosh, "Text summarization using Wikipedia," *Information Processing & Management*, vol. 50, pp. 443-461, 2016.

[11] Janjanam, P. and C.P. Reddy. "Text Summarization: An Essential Study," in 2019 I*nternational Conference on Computational Intelligence in Data Science (ICCIDS)*, 2019.

[12] Tandel, A., et al. "Multi-document text summarization - a survey," *in 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, 2016.

[13] McKeown, K., et al., "Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster," *Morgan Kaufmann Publishers Inc*, 2003.

[14] McKeown, K., et al." From text to speech summarization," *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing 2005*. vol. 5, pp. 997-v1000.

[15] Lei, K., et al., "Tag Recommendation by Text Classification with Attention-Based Capsule Network," *Neurocomputing*, pp.65-73, 2020.

[16] Zuo, Y., et al., "A Tag-aware Recommendation Algorithm Based on Deep Learning and Multi-objective Optimization," *in 2023 International Conference on Pattern Recognition, Machine Vision and Intelligent Algorithms (PRMVIA, pp.42-46*, 2023.

[17] Chatti, M.A., et al., "Tag-based collaborative filtering recommendation in personal learning environments," *IEEE Transactions on Learning Technologies*, vol. **6**, pp. 337-349, 2016.

[18] Shepitsen, A., et al., "Personalized recommendation in social tagging systems using hierarchical clustering," *RecSys'08: Proceedings of the 2008 ACM Conference on Recommender Systems*, 2008, pp. 259-266.

[19] Symeonidis, P., A. Nanopoulos, and Y. Manolopoulos, "Tag recommendations based on tensor dimensionality reduction," *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, 2008. pp. 43-50.

[20] Shang, M.-S., et al., "Collaborative filtering with diffusion-based similarity on tripartite graphs," *Physica A: Statistical Mechanics and its Applications*, vol. **389** pp. 1259-1264, 2010.

[21] Tang, J., et al., "Cross-domain collaboration recommendation," *in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, Association for Computing Machinery: Beijing, China. pp. 1285–1293.

[22] Zhang, X., J. Zhao, and Y. LeCun, "Character-level convolutional networks for text

[DOI: 10.61186/jsdp.21.4.15]

[34] Zhao, W., Z. Guan, and Z. Liu, "Ranking on heterogeneous manifolds for tag recommendation in social tagging services, " *Neurocomputing*, vol. **148**: pp. 521-534, 2015.

[35] Krestel, R., P. Fankhauser, and W. Nejdl, "Latent dirichlet allocation for tag recommendation," *in Proceedings of the third ACM conference on Recommender systems*, 2009, Association for Computing Machinery: New York, New York, USA. pp. 61–68.

[36] Weston, J., S. Chopra, and K. Adams, "#TagSpace: Semantic Embeddings from Hashtags," *Conference: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1822-1827.

[37] Ghabayen, A. and S.A. Mohd Noah, "Using Tags for Measuring the Semantic Similarity of Users to Enhance Collaborative Filtering Recommender Systems," I*nternational Journal on Advanced Science, Engineering and Information Technology (IJASEIT)*, vol **7**, pp. 2063-2070, 2017.

[۳۸] حمیدزاده، جواد، مرادی، منا، «بهبود پالایش مشارکتی در سیستم‌های توصیه‌گر به‌کمک خوشه‌بندی فازی C–میانگین مرتب شده و الگوریتم ازدحام ذرات تطبیقی – آشوبی»، *پردازش علائم و داده‌ها*، شمارهٔ ۱ (۵۹ پیاپی)، صص ۱۱۱-۱۲۲، ۱۴۰۳.

[۳۹] بحرانی، پیام، مینایی بیدگلی، بهروز، پروین، حمید، میرزا رضایی، میترا، و کشاورز، احمد، «ارائهٔ یک سامانهٔ پیشنهادگر محتوا-مشارکتی مبتنی بر خوشه‌بندی و هستان‌شناسی»، *پردازش علائم و داده‌ها*، شمارهٔ ۳ (۵۳ پیاپی)، صص ۱۴۷-۱۶۲، ۱۴۰۲.

[40] Pan, R., P. Dolog, and G. Xu., "KNN-Based Clustering for Improving Social Recommender Systems," i*n Agents and Data Mining Interaction*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[41] Gemmell, J., et al., "Adapting K-nearest neighbor for tag recommendation in Folksonomies," *in Proceedings of the 7th International Conference on Intelligent Techniques for Web Personalization & Recommender Systems* - Vol. 528. 2009, CEUR-WS.org: Pasadena, California. pp. 69–80.

[42] School of Software, X.U., Urumqi 830008, China, et al., "Extractive based Text Summarization Using KMeans and TF-IDF," *International Journal of Information Engineering and Electronic Business*, 2019. **11**(3): p. 33-44, 2019.

[43] Mahdi, A.E., A. Alahmadi, and A. Joorabchi, "Combining Bag-of-Words and Bag-of-Concepts Representations for Arabic Text Classification," *in 25th IET Irish Signals & Systems Conference 2014 and 2014 China-Ireland International Conference on Information and Communities Technologies (ISSC 2014/CIICT 2014).* 2014. Institution of Engineering and Technology.

[44] Weinberger, K., et al., "Feature hashing for large scale multitask learning," *in Proceedings* classification," *in Proceedings of the 28th International Conference on Neural Information Processing Systems* - Vol 1. 2015, MIT Press: Montreal, Canada. pp. 649–657.

[23] Li, H., et al., "Read, Watch, Listen, and Summarize: Multi-Modal Summarization for Asynchronous Text, Image, Audio and Video," *IEEE Transactions on Knowledge and Data Engineering*, 2019, vol. **31**, pp. 996-1009.

[24] Moradi, M. and N. Ghadiri, "Different approaches for identifying important concepts in probabilistic biomedical text summarization," *Artificial Intelligence in Medicine*, 2018, vol. **84**, pp. 101-116, 2018.

[25] Rautray, R. and R.C. Balabantaray, "An evolutionary framework for multi document summarization using Cuckoo search approach," *MDSCSA. Applied Computing and Informatics*, vol. **14**, pp. 134-144, 2018.

[26] Rautray, R. and R.C. Balabantaray, "Comparative Study of DE and PSO over Document Summarization," in I*ntelligent Computing, Communication and Devices, L.C. Jain, S .Patnaik, and N. Ichalkaranje, Editors*. *Springer India: New Delhi.* pp. 371-377, 2015.

[27] Nasr Azadani, M., N. Ghadiri, and E. Davoodijam, "Graph-based biomedical text summarization: An itemset mining and sentence clustering approach," J*ournal of Biomedical Informatics*, vol. 84: pp. 42-58, 2018.

[28] Baralis, E., et al., "MWI-Sum: A Multilingual Summarizer Based on Frequent Weighted Itemsets," *ACM Transactions on Information Systems*, vol. 34**,** pp. 1-35, 2015.

[29] Tohalino, J.V. and D.R. Amancio. "Extractive Multi-document Summarization Using Dynamical Measurements of Complex Networks," *in 2017 Brazilian Conference on Intelligent Systems (BRACIS)*, 2017.

[30] Afsharizadeh, M., H. Ebrahimpour-Komleh, and A. Bagheri. "Query-oriented text summarization using sentence extraction technique," *in 2018 4th International Conference on Web Research (ICWR)*. 2018.

[31] Lei, K., et al., "Tag recommendation by text classification with attention-based capsule network," *Neurocomputing*, vol. 391, pp. 65-73, 2020.

[32] Feng, W. and J. Wang, "Incorporating heterogeneous information for personalized tag recommendation in social tagging systems," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2012, *Association for Computing Machinery: Beijing, China*, pp. 1276–1284.

[33] Fang, X., et al., "Personalized tag recommendation through nonlinear tensor factorization using gaussian kernel," *in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, AAAI Press: Austin, Texas. pp. 439–445.

[DOI: 10.61186/jsdp.21.4.15]

of the 26th Annual International Conference on Machine Learning*. 2009, Association for Computing Machinery: Montreal, Quebec, Canada. pp. 1113–1120.

[45] Li, P., C. Tang, and X. Xu, "Video summarization with a graph convolutional attention network," *Frontiers of Information Technology & Electronic Engineering*," vol. 22(6): pp. 902-913, 2021.

[46] Zhang, K., et al. "Video Summarization with Long Short-Term Memory," i*n Computer Vision – ECCV 2016*, 2016. Cham: Springer International Publishing.

[47] ji, Z., et al., "Video Summarization With Attention-Based Encoder–Decoder Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017. PP 208-214.

[48] Zhao, B., X. Li, and X. Lu, "Property-Constrained Dual Learning for Video Summarization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31(10): pp. 3989-4000, 2020

[49] Zhang, P. and Z. Yang, "A Novel AdaBoost Framework With Robust Threshold and Structural Optimization," *IEEE Transactions on Cybernetics*, vol. 48(1): p. 64-76., 2018

[50] Wang, Y., et al., "Bernoulli random forests: closing the gap between theoretical consistency and empirical soundness," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 2016, AAAI Press: New York, New York, USA. pp. 2167–2173.

[51] Friedman, J.H., "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 4, pp. 367-378, 2002.

[52] Cho, K., et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, vol. 1, pp.1724-1734.

[53] Zhang, S., et al. "Bidirectional Long Short-Term Memory Networks for Relation Classification," in *PACLIC*, 2015., pp. 73-78.

[54] Zhou, P., et al., "Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification," *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016. pp. 207-212.

[55] Kim, Y., "Convolutional Neural Networks for Sentence Classification," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

[56] Yin, W., et al., "ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs," *Transactions of the Association for Computational Linguistics*, vol. 4, 2015.

[57] Schwenk, H., et al. "Very Deep Convolutional Networks for Text Classification," *in EACL*, 2017.

[58] Sabour, S., N. Frosst, and G.E. Hinton, "Dynamic routing between capsules," i*n Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, Curran Associates Inc .:Long Beach, California, USA. pp. 3859–3869.

[59] Yang, M., et al., "Investigating Capsule Networks with Dynamic Routing for Text Classification," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3110-3119.

مهسا رحیمی رسکتی دکترای مهندسی نرم‌افزار، استادیارمعلم آموزش و پرورش استان مازندران هستند. زمینه‌های پژوهشی مورد علاقهٔ ایشان مهندسی نرم‌افزار، کاوش داده، خلاصه‌سازی و رمزگذاری داده است.

نشانی رایانامهٔ ایشان عبارت‌است از:

mr2.mco@gmail.com

همایون موتمنی دکترای مهندسی کامپیوتر و دارای مرتبهٔ علمی استادی و عضو هیئت علمی دانشگاه آزاد اسلامی واحد ساری هستند. زمینه‌های پژوهشی مورد علاقهٔ ایشان مهندسی نرم‌افزار و پتری‌نت است.

نشانی رایانامهٔ ایشان عبارت‌است از:

h_motameni@yahoo.com

ابراهیم اکبری دکترای مهندسی کامپیوتر، دانشیار و عضو هیئت علمی دانشگاه آزاد اسلامی ساری هستند. زمینه‌های پژوهشی مورد علاقهٔ ایشان مباحثی نظیر الگوریتم‌های داده‌کاوی، بهینه‌سازی، طبقه‌بندی و خوشه‌بندی داده‌ها است.

نشانی رایانامهٔ ایشان عبارت‌است از:

akbari@iausari.ac.ir

حسین نعمت‌زاده دکتری مهندسی کامپیوتر، استادیار و عضو هیئت علمی دانشگاه آزاد اسلامی واحد ساری هستند. زمینه‌های پژوهشی مورد علاقهٔ ایشان مباحثی نظیر الگوریتم‌های داده‌کاوی، بهینه‌سازی و یادگیری ماشین است.

نشانی رایانامهٔ ایشان عبارت است از:

hn_61@yahoo.com

[DOI: 10.61186/jsdp.21.4.15]