ارائهٔ روشی برای استخراج کلمات کلیدی و وزن‌دهی کلمات برای بهبود طبقه‌بندی متون فارسی

وحیده رضایی[۱]، مجید محمدپور[۲]، حمید پروین[۳] و صمد نجاتیان*[۴]

[۱] گروه ریاضی، واحد یاسوج، دانشگاه آزاد اسلامی یاسوج، ایران

[۲] گروه مهندسی کامپیوتر، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

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

[۴] گروه مهندسی برق، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

[۱، ۲ و ۴] باشگاه پژوهشگران و نخبگان، واحد یاسوج، دانشگاه آزاد اسلامی، یاسوج، ایران

[۳] باشگاه پژوهشگران و نخبگان، واحد نورآباد ممسنی، دانشگاه آزاد اسلامی، نورآباد ممسنی، ایران

چکیده

با توجه به گسترش روزافزون اطلاعات و وجود حجم انبوه متون غیرساخت یافته، استفاده از کلمات کلیدی نقش مهمی در بازیابی اطلاعات دارد. این درحالی است که استخراج کلمات کلیدی به‌صورت دستی مشکلات زیادی دارد. بنابرین استخراج کلمات کلیدی به‌صورت خودکار از نیازهای ضروری فناوری امروزه است. در این پژوهش سعی شده با استفاده از اصطلاح‌نامه که از نظامی ساختارمند برخوردار است، کلمات کلیدی بامعناتری از متون استخراج کرد و با آن‌ها طبقه‌بندی متون فارسی را بهبود بخشید. مراحلی که برای افزایش جامعیت جستجو باید سپری شود به این صورت است که در مرحله نخست کلمات زائد حذف و باقی کلمات ریشه‌یابی می‌شود؛ سپس به‌کمک اصطلاح‌نامه کلمات هم‌معنی، اعم‌ها و اخص‌ها و همچنین وابسته‌ها پیدا و در ادامه برای مشخص‌شدن اهمیت نسبی کلمات یک وزن عددی به هر کلمه منسوب می‌شود که بیان‌گر میزان تأثیر کلمه در ارتباط با موضوع متن و درمقایسه با سایر کلمات به‌کاررفته در متن است. با توجه به مراحل بالا و به‌کمک اصطلاح‌نامه، طبقه‌بندی‌متون دقیق‌تر انجام می‌گیرد. در این روش از الگوریتم نزدیکترین همسایه (KNN) برای طبقه‌بندی استفاده می‌شود. الگوریتم KNN به‌خاطر سادگی و مؤثربودن آن در طبقه‌بندی متون بسیار به‌کار برده می‌شود. مبنای کار این الگوریتم، مقایسه متن آزمایش داده‌شده با متون آموزشی داده‌شده و به‌دست‌آوردن میزان شباهت بین آن‌ها است. نتایج آزمایش‌ها برروی چندین متن در موضوع‌های مختلف، نشان‌دهندهٔ دقت و توانایی روش پیشنهادی در استخراج کلمات کلیدی منطبق با خواست کاربر و درنتیجه طبقه‌بندی دقیق‌تر متون است.

واژگان کلیدی: اصطلاح‌نامه، بازیابی اطلاعات، استخراج کلمات کلیدی، وزن‌دهی.

An Approach for Extraction of Keywords and Weighting Words for Improvement Farsi Documents Classification

Vahideh Rezaei [1], Majid Mohammadpour [2], Hamid Parvin [3] & Samad Nejatian* [4]

[1] Department of Mathematics, Yasooj Branch, Islamic Azad University, Yadooj, Iran.

[2] Department of Computer Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran.

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

* Corresponding author

* نویسندهٔ عهده‌دار مکاتبات

[4] Department of Electrical Engineering, Yasooj Branch, Islamic Azad University, Yasooj, Iran

[1, 2, 4] Young Researchers and Elite Club, Yasooj Branch, Islamic Azad University, Yasooj, Iran

[3] Young Researchers and Elite Club, Nourabad Mamasani Branch, Islamic Azad University, Nourabad Mamasani, Iran

Abstract

Due to ever-increasing information expansion and existing huge amount of unstructured documents, usage of keywords plays a very important role in information retrieval. Because of a manually-extraction of keywords faces various challenges, their automated extraction seems inevitable. In this research, it has been tried to use a thesaurus, (a structured word-net) to automatically extract them. Authors claim that extraction of more meaningful keywords out of documents can be attained via employment of a thesaurus. The keywords extracted by applying thesaurus, can improve the document classification. The steps to be taken to increase the comprehensiveness of search should be such that in the first step the stop words are removed and the remaining words are stemmed. Then, with the help of a thesaurus are found words equivalent, hierarchical and dependent. Then, to determine the relative importance of words, a numerical weight is assigned to each word, which represents effect of the word on the subject matter and in comparison with other words used in the text. According to the steps above and with the help of a thesaurus, an accurate text classification is performed. In this method, the KNN algorithm is used for the classification. Due to the simplicity and effectiveness of this algorithm (KNN), there is a great deal of use in the classification of texts. The cornerstone of KNN is to compare with the text trained and text tested to determine their similarity between. The empirical results show the quality and accuracy of extracted keywords are satisfiable for users. They also confirm that the document classification has been enhanced. In this research, it has been tried to extract more meaningful keywords out of texts using thesaurus (which is a structured word-net) rather than not using it.

Keywords: thesaurus, information retrieval, extraction of keywords, weight.

۱- مقدمه

طبقه‌بندی متون با استفاده از شاخص‌های کلیدی نقش مهمی در بازیابی اطلاعات دارد. امروزه در بیش‌تر کتب و مقالات به‌خصوص در زمینهٔ علمی و فنی با بررسی شاخص به مطالب و موضوعات مطرح‌شده می‌توان پی برد. ساخت شاخص برای متن‌ها، استفاده از آن‌ها را برای پژوهش‌گران و خوانندگان آسان می‌سازد. یکی از مشکل‌ترین وظایف سامانه‌های مدیریت متن، بازیابی متن و امکان جستجوی کارا روی اطلاعات متنی است. بنابراین در یک سامانه بازیابی متن، لازم است ابتدا سامانه، مجموعه اقداماتی را در جهت ساخت یک شاخص مناسب، مجهز، و کارآمد روی واژه‌های متن انجام دهد. پس از ساخت شاخص، سامانه با استفاده از آن، در جواب به پرس و جوی کاربر، متونی را که مربوط به واژه‌های مورد درخواست کاربر هستند، یافته و می‌تواند ارائه کند.

اصولاً هر متنی دارای دو بخش است: بخش نخست شامل اطلاعاتی است که نسبت به محتوای متن، خارجی محسوب می‌شوند؛ مثل نام نویسنده، تاریخ و محل انتشار و نام منتشرکننده، بخش دوم شامل محتویات اصلی متن است. در کتابخانه‌ها بخش نخست با نام دسته‌بندی توصیفی[1] و بخش دوم با نام دسته‌بندی موضوعی[2] شناخته می‌شود. شاخص‌گذاری به عملیات شناخت محتوای متن گفته می‌شود، و هنگامی که این عملیات به‌کمک وسایل پیشرفته رایانه‌ای صورت پذیرد، شاخص‌گذاری خودکار نامیده می‌شود [28].

در سیستم‌های بازیابی متن، شاخص می‌تواند به‌طورکامل به‌صورت خودکار تولید شود. تحقیقات درباره ایجاد و یا بهبود روش‌های تولید خودکار شاخص و نیز جستجوی اطلاعات در متون برای زبان‌های مختلف همواره در جریان بوده است. حساس‌ترین و مشکل‌ترین مرحله‌ای که

[1] descriptive cataloging

[2] subject cataloging

در روند شاخص‌گذاری خودکار باید طی شود، انتخاب واژه‌هایی است که برای ساخت شاخص به‌کار می‌روند.

درعمل، شاخص‌گذاری روی تمام واژه‌های متن، سربار بسیار زیاد دارد؛ ضمن این‌که شاخص‌گذاری روی تمام واژه‌ها یک کار غیرضروری است. کافی است تنها واژه‌هایی در شاخص به‌کار روند که نشان‌دهندهٔ محتوای متن مربوطه هستند. در واقع واژه‌هایی که مورد علاقه کاربر بوده، و توسط وی جستجو می‌شوند.

در زمینهٔ پردازش اطلاعات، سامانه‌های بسیاری ایجاد شده‌اند. این سامانه‌ها در پنج گروه عمده دسته‌بندی می‌شوند، که عبارتند از: سامانه‌های اطلاعات مدیریتی[1]، سامانه‌های مدیریت پایگاه داده[2]، سامانه‌های تصمیم[3]، سامانه‌های پرسش و پاسخ[4] و سامانه‌های بازیابی اطلاعات[5] [33]. سامانه‌های بازیابی متن در گروه سامانه‌های بازیابی اطلاعات قرار می‌گیرند. از آن‌جا که شباهت‌های زیادی بین سامانه‌های بازیابی اطلاعات و سامانه‌های مدیریت پایگاه داده وجود دارد، گاهی برخی افراد این دو سامانه را به‌اشتباه یکی در نظر می‌گیرند. اکنون به بررسی خصوصیات این دو سامانه می‌پردازیم:

وظیفه یک سامانه مدیریت پایگاه داده، ذخیره‌سازی، نگهداری و بازیابی دقیق داده‌های موجود در سامانه است. اطلاعات به‌شکل عناصر داده‌ای معینی است که در جداول ذخیره می‌شوند. هر درخواست جستجو باید بیان‌گر مقادیر معینی از مشخصه‌های رکورد باشد. خروجی در این‌گونه سامانه‌ها، رکوردهای مجزا، بخش‌هایی از رکوردها یا جداول و ... است.

در سامانه‌های بازیابی اطلاعات، عملیات پردازش اطلاعات بر روی مستند[6] انجام می‌گیرد و وظیفه سامانه بازیابی، ذخیره، ارائه و ایجاد امکان دسترسی به مستندات یا نماینده آن‌ها است. در سامانه‌های بازیابی متن، اطلاعات ورودی به‌صورت متن زبان طبیعی است (متن کامل و یا گلچین یا چکیده متن کامل). خروجی سامانه‌های بازیابی اطلاعات در پاسخ به یک درخواست جستجو، به شکل مجموعه‌ای از ارجاعات است. این ارجاعات به کاربران سامانه، اطلاعاتی در مورد بخش‌های مورد علاقه آن‌ها می‌دهد. کاربران سامانه‌های بازیابی اطلاعات، نیازهای اطلاعاتی بسیار گسترده و متفاوتی دارند. این کاربران ممکن است، پژوهش‌گرانی باشند که به دنبال مقاله در مورد موضوع خاصی می‌گردند؛ و یا مهندسانی که می‌خواهند بدانند آیا ایده مورد نظر آن‌ها قبلاً به ثبت رسیده است؛ یا خریداران کالایی خاص باشند که اطلاعات جدیدی در مورد آن کالا می‌خواهند پیدا کنند؛ و یا وکلایی که به‌دنبال قوانین، رویه‌ها، و احکام صادره محاکم قضایی هستند. به‌عبارت دیگر، کاربران سامانه‌های بازیابی اطلاعات، در زمینه‌های مختلفی کار می‌کنند و دلایل متفاوتی، آن‌ها را به استفاده از امکانات سامانه‌های بازیابی اطلاعات، مجبور می‌کند [7].

مزایای استفاده از شاخص‌ها و عبارت‌های کلیدی :

۱- استخراج خودکار عبارت‌های کلیدی، یک متن بلند را به خلاصه‌ای کوتاه تبدیل می‌کند.

۲- عبارت‌های کلیدی به‌عنوان قسمتی از نتایج جستجو، همراه با سایر مشخصه‌های متن، بازیابی‌شده (همانند عنوان، قسمت هایی از متن، URL و ...) یا به‌جای آن‌ها می‌توانند نمایش داده شوند.

۳- در مواردی که به مشخصه‌هایی بیش از نام‌گذاری صرف، به‌منظور درک سریع‌تر متن نیاز داریم، شاخص‌ها می‌توانند مفید باشند. به‌عنوان مثال، اگر نام یک فایل یا نامه الکترونیکی به‌عنوان برچسب با شاخص ادغام شوند، حالت بهتری را ایجاد می‌کنند. در این حالت، مشاهده عبارت‌های کلیدی همراه با عنوان، به فهم محتوای نامه کمک بیشتری می‌کند.

۴- برجسته‌کردن شاخص‌ها در متون الکترونیکی به مرور سریع و اجمالی متن می‌تواند کمک کند.

۵- کمک به نویسنده یا ویراستار در تخصیص شاخص‌های کلیدی به متن. انجام این کار به‌صورت خودکار به‌عنوان یک استاندارد، نوعی یک‌دستی و مطابقت نوشته با کارکرد سامانه بازیابی اطلاعات و در نتیجه اطلاع‌رسانی صحیح‌تر را به همراه می‌تواند داشته باشد.

۶- در مواردی که با مشکل پهنای خط یا مطابق با اصول نمایش گرافیکی اطلاعات با محدودیت فضای نمایشی مواجه هستیم، نمایش عبارت‌های کلیدی بسیار مفید است.

---

[1] Management Information System

[2] Data Base Management System

[3] Decision Support System

[4] Question Answering System

[5] Information Retrieval

[6] document

۷- استخراج خودکار عبارت‌های نمایه‌ای متون نشریات و صفحات وب، خواندن و جستجوی اطلاعات نشریات را برای خوانندگان تسهیل می‌کند.

۸- حضور شاخص‌های کلیدی در نتایج جستجو به اصلاح و تعریف مجدد فرمول جستجو و حتی تغییر دیدگاه کاربران از ساختار موجود در یک زمینه خاص می‌تواند کمک کند؛ یعنی کاربران با افزودن یا حذف واژگان، دامنه جستجو را محدودتر کرده و ضریب دقت را می‌توانند بالاتر ببرند؛ درنتیجه، به بالابردن ضریب دقت یا گسترده‌ترکردن دامنهٔ جستجو و درنتیجه، بالابردن ضریب بازیابی کمک می‌کند. بنابراین می‌توان عبارت‌های کلیدی را به‌عنوان جزئی لازم برای سامانه‌های بازیابی و طبقه‌بندی اطلاعات معرفی کرد.

۹- در مفاهیم سازمان‌دهی اطلاعات در سامانه‌های بازیابی و طبقه‌بندی اطلاعات به‌گونه‌ای مؤثر از شاخص‌های کلیدی در دسته‌بندی و طبقه‌بندی مدارک می‌توان استفاده کرد.

در الگوریتمی که در این پژوهش مطرح می‌شود، ضمن به‌کارگیری روابط اعم و اخص و هم‌چنین روابط هم‌خانوادگی میان لغات، میزان دقت در طبقه‌بندی متون فار سی بهبود می‌یابد.

۱-۱- هدف و ضرورت انجام پژوهش

هنوز هم پراستفاده‌ترین و حجیم‌ترین اطلاعات موجود، متون غیر ساخت‌یافته هستند. به‌عنوان مثال در سامانه‌های بایگانی، اطلاعات متنی زیادی در قالب نامه‌ها، دستورالعمل‌ها، جزوات و ... وجود دارد، و یا در یک کتابخانه رقمی، غالب اطلاعات به‌صورت متن است. سامانه نگهداری و به‌روزرسانی قوانین (حقوقی، مالیاتی، جزایی، و ...)، سامانه نگهداری و بازیابی مقالات و آرشیو الکترونیکی روزنامه‌ها و مجلات، سامانه‌های تولید و نگهداری اطلاعات وب (www)، و نیز سامانه‌های بازیابی رایانامه[1] نمونه‌هایی دیگر از این دست هستند. کلمات کلیدی درحقیقت نقش کلیدی در بازیابی اطلاعات دارد. امروزه در بیش‌تر کتب و مقالات به‌خصوص در زمینهٔ علمی و فنی به مطالب و موضوعات مطرح شده با بررسی شاخص می‌توان پی برد. ساخت شاخص برای متن‌ها، استفاده از آن‌ها را برای پژوهش‌گران و خوانندگان آسان می‌سازد.

یکی از مشکل‌ترین وظایف سامانه‌های مدیریت متن، بازیابی متن[1] و امکان جستجوی کارا روی اطلاعات متنی است. بنابراین در یک سامانه بازیابی متن، لازم است تا ابتدا سامانه، مجموعه اقداماتی را در جهت ساخت یک شاخص مناسب، مجهز و کارآمد روی واژه‌های متن انجام دهد. پس از ساخت شاخص، سامانه می‌تواند با استفاده از آن، در جواب به پرس‌وجوی کاربر، متونی را که مربوط به واژه‌های مورد درخواست کاربر هستند، یافته و ارائه کند.

از آن‌جاکه تعداد مستندات الکترونیکی فارسی به‌سرعت رو به رشد است، به‌کارگیری روش‌های کارامد جهت طبقه‌بندی اطلاعات فارسی بسیار مورد اهمیت است. کلمات کلیدی مجموعه‌ای از لغات مهم در یک مستند هستند که توصیفی از محتوای مستند را فراهم می‌آورند و برای اهداف مختلفی قابل استفاده هستند. به‌عنوان مثال در مقالات علمی برای این‌که خواننده بتواند درک مناسبی از مقاله داشته باشد، شاخص‌ها و کلمات کلیدی مقاله عنوان می‌شوند. استخراج شاخص‌ها از مستندات، یک عملیات مهم در فرآیندهایی مانند طبقه‌بندی و یا استخراج اطلاعات معنایی است. هم‌چنین شاخص‌ها می‌توانند در موتورهای جستجو به‌منظور برگرداندن نتایج دقیق‌تر و در زمان کوتاه‌تر مورد استفاده قرار گیرند. شاخص‌ها نکات اصلی متن را توصیف می‌کنند؛ لذا می‌توانند به‌عنوان یک ابزار، برای اندازه‌گیری شباهت متون مختلف به‌منظورطبقه‌بندی متون مورد استفاده قرار گیرند. درمجموع شاخص‌ها ابزار مفیدی برای جستجوی حجم زیادی از مستندات در زمان کوتاه هستند. با وجود اهمیت سرشار، تعداد بسیار کمی از مستندات در حال حاضر حاوی شاخص هستند. درواقع بسیاری از نویسندگان تمایلی به استخراج شاخص ندارند و تنها درصورتی‌که مجبور به استخراج آن‌ها باشند، آن‌ها را در مستندات قرار می‌دهند. استخراج شاخص به‌طور دستی فرایندی بسیار دشوار و زمان‌بر است؛ بنابراین نیاز به یک فرایند خودکار است که آن‌ها را از مستندات خارج کند.

جستجو بر اساس شاخص طبقه‌بندی، یک ابزار قابل استفاده و قدرتمند است که امکان جستجوی سریع اطلاعات را در مجموعه های بزرگ از اسناد به‌راحتی میسر می‌سازد. این امر کاربر را در فراگیری قواعد و دستور یک زبان پرس‌وجوی ساخت‌یافته، به‌عنوان مثال پرس‌وجوی بولی، SQL و یا $XQuery$ و درک معانی پیچیده آن‌ها آزاد می‌گذارد.

[1]-Email

طبقه‌بندی متون براساس شاخص، یک مسأله بسیار مهم در پردازش زبان فارسی است. در زبان فارسی کلمات صورت‌های نگارشی پیچیده‌ای دارند و پوشش کلیه حالات دستوری کلمات با به‌کارگیری یک سری قواعد معین، ناممکن است؛ به‌همین دلیل استخراج شاخص و طبقه‌بندی متون به‌طور خودکار از در مستندات فارسی دشوار و پیچیده است. از طرفی کارهای چندانی در زمینهٔ استخراج شاخص در متون فارسی انجام نشده است. بدون استخراج شاخص، بسیاری از کاربردهای بازیابی اطلاعات مانند جستجوی متن، طبقه‌بندی مستندات، پالایش اطلاعات و خلاصه‌سازی متن، به نتایج مطلوبی نمی‌توانند دست یابند. بنابراین در این مقاله یک روش طبقه‌بندی متون فارسی ارایه می‌شود؛ به‌گونه‌ای که کارایی روش‌های بازیابی و طبقه‌بندی اطلاعات را افزایش دهند.

در این مقاله سعی شده است تا با استفاده از روابط موجود بین کلمات به‌کمک اصطلاح‌نامه روش مناسبی برای ساخت خودکار شاخص در متون فارسی ارائه شود؛ ضمن این‌که با به‌کارگیری روش طبقه‌بندی KNN ، متون موجود با دقت بر اساس محتوا و موضوع دسته‌بندی شوند.

هدف از این مقاله ارائهٔ راه‌کاری جدید برای وزن‌دهی کلمات به‌کمک اصطلاح‌نامه در طبقه‌بندی متون فارسی جهت افزایش جامعیت جستجو است. با توجه به رشد و گسترش حجم اطلاعات و به‌موازات آن ضرورت به‌کارگیری و استفاده مؤثر از منابع اطلاعاتی، یکی از مهم‌ترین و اساسی‌ظترین نیازهای موجود، قابلیت دست‌یابی به اطلاعات مورد نیاز در مدت زمان مناسب است. درواقع انجام جستجو و یافتن اطلاعات مورد نظر بر اساس خواسته‌های کاربر اهمیت بالایی دارد.

درواقع در این پژوهش یک الگوریتم مناسب، جهت طبقه‌بندی متون مختلف بر اساس مفاهیم روابط اخص اعم و هم‌خانواده‌بودن میان کلمات استفاده‌شده در متون ارائه شده است. استفاده از این روابط موجب می‌شود پس از جداسازی واژه‌های یک متن، دقت فرآیند وزن‌دهی به کلمات افزایش یابد و طبقه‌بندی متون با ضریب اطمینان بیشتری صورت گیرد.

از آن‌جا که در این راستا پژوهش‌های زیادی در جهت کشف روابط مختلف بین کلمات در متون مختلف فارسی صورت گرفته، پژوهش‌گران این امر هر یک به‌نوعی نسبت به ثبت یافته‌های خود و روابط تعریف‌شده لغات با یکدیگر، در فرهنگ اصطلاحات زبان فارسی یا همان "اصطلاح‌نامه" اقدام کرده‌اند و هر چه این اصطلاح‌نامه به‌مرور زمان تکمیل‌تر شود تجزیه و تحلیل متون فارسی با سرعت، دقت و اطمینان بیشتری در آینده مواجه خواهد بود.

۲-۱- تحقیق ساختار پژوهش

در این بخش ساختار مقاله و بخش‌های مربوطه به‌اجمال توصیف می‌شود. این مقاله شامل پنج بخش است: بخش نخست، مقدمه و کلیات را بیان می‌کند. در بخش دوم مروری بر کارهای گذشته صورت گرفته است. فصل سوم روش پیشنهادی را تشریح می‌کند. فصل چهارم روش پیشنهادی را مورد ارزیابی و آزمایش قرار می‌دهد. فصل پنجم به ارائه نتایج حاصل از تحقیق می‌پردازد و درنهایت پیشنهادهایی درخصوص کارهای آینده ارائه می‌شود.

۲- ادبیات پژوهش و کارهای مرتبط پیشین

در زمینهٔ طبقه‌بندی متون، اصطلاحات و واژگانی وجود دارند که برخی از آن‌ها به شرح جدول (۱) اعلام می‌شود [2].

(جدول-۱): اصطلاحات و واژگان
(Table-1): Terminology and vocabulary

اصطلاح‌نامه[1]	مجموعه‌ای است که در آن تنظیم و مرتب کردن واژه‌ها و عبارت‌های زبان، نه بر حسب الفبا، بلکه برحسب مفاهیمی است که بیان می‌کنند فرهنگ مفاهیم با فرهنگ لغات متفاوت است.
پیکره زبانی[2]	مجموعه‌ای از نوشتار در یک زبان است که می‌توان ویژگی‌های زبان را با استفاده از آن بازنمایی نمود.
واژگان[3]	مجموعه‌ای از لغات و کلیه مشتقات آن است که در بعضی اوقات قوانین تولید واژه را نیز شامل می‌شود.
مجموعه آموزشی	مجموعه اطلاعاتی است که برای آموزش الگوریتم‌های با ناظر استفاده می گردند.
مجموعه آزمایشی	مجموعه اطلاعاتی است که برای آزمایش الگوریتم‌های با ناظر و یا بدون ناظر استفاده می‌گردند.
کلمات بی‌ارزش	کلماتی هستند که هیچ‌گونه ارزش معنایی و یا مفهومی از نقطه نظر طبقه‌بندی ندارند.
کلمه کلیدی	کلمات خاصی که یک متن می‌تواند بر مبنای آن

[1] Thesaurus
[2] Corpus
[3] Lexicon

	اندیس‌گذاری گردد. این کلمات یا عبارات به نوعی متن را طبقه‌بندی می‌کنند.
عبارت	دو یا چند کلمه که ترکیب مشخصی از آن‌ها مفهوم خاصی را منتقل می‌کند. برای مثال" سازمان" + "ملل " و از طرف دیگر" سازمان ملل"
شاخص	شاخص‌ها مجموعه‌ای منظم از کلمات نشانه‌گذاری شده می‌باشد تا کاربران را قادر سازد اطلاعاتی که محل آن‌ها در مدرک مشخص‌شده پیدا کنند.

۱-۲- طبقه‌بندی متون

درصورتی‌که مجموعه‌ای از متون $D = \{(d_1, y_1), \ldots, (d_i, y_i), \ldots, (d_n, y_n)\}$ داشته باشیم، به‌طوری‌که n تعداد متون و $d_i = [w_i, 1, \ldots, w_i, k, \ldots, w_i, |d_i|]$ ، متن iام این مجموعه باشد wi, k کلمه kام متن iام باشد و y_i به طبقه‌ای که متن به آن متعلق است (یعنی $y_i \in C$ به‌طوری‌که $C = \{c_1, c_2, \ldots, c|C|\}$ مجموعه طبقه‌های از پیش تعریف‌شده در سامانه باشد) اشاره کند. هدف در طبقه‌بندی متون استنتاج یک تابع رابطه‌ای f است به‌نحوی که $y_i = f(d_i)$ باشد. یا به‌صورت کامل‌تر، طبقه‌بندی متون تعیین یک مقدار بولی[1] برای هر جفت $< d_j, c_i > \in \hat{D} * C$ در جایی که D مجموعه‌ای از متون و C مجموعه طبقه‌های از پیش تعیین‌شده است. مقدار T تعیین می‌کند که متن d_j به طبقه c_i متعلق است و مقدار F نیز عدم تعلق متن d_j به c_i را نشان می‌دهد. هدف این‌جا به‌دست‌آوردن تخمین تابع $\emptyset: D * C \to \{T, F\}$ است. از این قسمت به بعد فرض می‌شود (دانشگاه علم و صنعت ایران ۱۳۸۸) [2]:

- طبقه‌ها تنها برچسب‌های نمادین هستند و هیچ دانش اضافی (به‌لحاظ اجرایی یا تعریفی) با خود به‌همراه ندارند.
- دانش از بیرون[2] (اطلاعاتی از منبع خارجی جهت طبقه‌بندی) موجود نباشد. بنابراین طبقه‌بندی می‌بایست تنها بر اساس دانش از درون[3] (دانشی که از خود متن به دست می‌آید) انجام گیرد. به عبارت دیگر اطلاعات دیگر هم‌چون نویسنده، تاریخ انتشار در دسترس نباشد.

روش‌های طبقه‌بندی متون که در مورد آن‌ها بحث خواهد شد، به‌طورکامل کلی بوده و برای یک زمینه خاص نمی‌باشد. درحقیقت این پیش‌فرض‌ها اگرچه هزینه‌های طبقه‌بندی را افزایش می‌دهد، ولی برای قانونی‌بودن عملیات طبقه‌بندی متون اجباری است [12]. طبقه‌بندی بر مبنای دانش از درون، یعنی یک متن تنها بر اساس اطلاعات معنایی‌اش طبقه‌بندی می‌شود. در زمینه طبقه‌بندی متون، کارهای زیادی انجام گرفته که در ادامه به برخی از آن‌ها اشاره شده است.

در سال ۱۳۹۵، راد و همکارانش، یک روشی جدید را برای شاخص‌گذاری خودکار و استخراج کلمات کلیدی برای بازیابی اطلاعات و خوشه‌بندی متون ارائه داده است [1]. در این مقاله سعی شده است با استفاده از اطلاعات زبان‌شناختی و اصطلاح‌نامه، کلمات کلیدی بامعناتری ارائه شود. با استفاده از اصطلاح‌نامه که از نظامی ساختارمند برخوردار است شبکه کلمات کلیدی، شامل کلمات هم‌ارز، کلمات سلسله‌مراتبی و وابسته را می‌توان تکمیل کرده و افزایش داد. بنابراین توافق بین جستجوی کاربران و کلمات کلیدی متنی را بیشتر می‌توان کرد و جامعیت جستجو را افزایش داد. در مرحله نخست کلمات غیرمهم و عمومی حذف می‌شوند؛ سپس کلمات متن ریشه‌یابی می‌شوند و در ادامه برای مشخص‌شدن اهمیت نسبی کلمات با استفاده از روش‌های وزن‌دهی یک وزن عددی به هر کلمه منسوب می‌شود که بیانگر میزان تأثیر کلمه در ارتباط با موضوع متن و در مقایسه با سایر کلمات به‌کاررفته در متن است. مجموعه عملیات بالا به‌خصوص استفاده از اصطلاح‌نامه باعث می‌شود که دسته‌بندی متون دقیق‌تر انجام گیرد و به‌نوعی رده علمی سلسله‌مراتبی متون در حوزه بازیابی اطلاعات نیز مشخص می‌شود[1].

در سال ۲۰۰۴، دیگان و هم‌کارانش از دو منبع مختلف گنج‌واژه و صفحات خبری وب، برای استخراج کلمات استفاده کردند. آن‌ها برای هر لغت از مدل‌های احتمالی مارکوف و بین هر لغت و اصطلاح‌نامه از برچسب‌های معنایی استفاده کردند. در کار آن‌ها استخراج لغات کلیدی خروجی یک فرآیند پیچیده بود. دقت این سامانه بسیار قابل توجه بود و این نخستین سامانه‌ای بود که از گنج‌واژه استفاده می‌کرد [13].

در سال ۲۰۰۶، هیون سعی کرد از رابطه سلسله‌مراتبی و طبقه‌بندی علوم در استخراج کلمات کلیدی استفاده کند. مزیت این روش، استخراج ک کلمات کلیدی و محدودیت آن، این بود که فقط به بخشی از خواص گنج‌واژه که همان خاصیت طبقه‌بندی است، پرداخته شده بود [18].

[1] Boolean

[2] Exogenous Knowledge

[3] Endogenous Knowledge

در سال ۲۰۰۶، ویتن و مدلی سعی کردند که روشی جدید و پیشرفته برای استخراج کلمات کلیدی بر اساس اطلاعات معنایی پیشنهاد تا مشکلات موجود پیشین را حل کنند. آن‌ها یک الگوریتم بر اساس اندیس‌های گنج‌واژه برای مستندات که KE + + نامیده می‌شد، ارائه دادند که در آن از روش‌های فنی یادگیری ماشین و اطلاعات معنایی استفاده می‌شد. مزیت اصلی این سامانه فرهنگ لغات کنترل‌شده است[38].

مزدک و هسل یک سامانه به‌نام FarsiSum [20] ارائه دادند که نسخه تغییر یافته یک سیستم خلاصه‌سازی متون سوئدی به‌نام Swe Sum برای پوشش زبان فارسی است[14].

زمانی‌فر، مینایی و شریفی نیز یک سیستم خلاصه سازی برای متون فارسی با استفاده از استخراج کلمات کلیدی ارائه کردند. هم‌چنین آن‌ها به توسعه این سامانه پرداخته و راه‌کاری برای کشف مشابهت متون و بازیابی اطلاعات ارائه کرده‌اند [37].

آقای پروین و همکاران در مقاله خود روشی نوین برای بهبود طبقه‌بندی متون فارسی ارائه داده‌اند که در آن، از اصطلاح‌نامه به‌عنوان ابزاری سودمند برای دست‌یابی به تکرار کلمات در متون فارسی استفاده می‌شود. در اصطلاح‌نامه مورد استفاده، سه نوع رابطه در نظر گرفته شده که این نخستین تلاش برای به‌کارگیری اصطلاح‌نامه فارسی در زمینه بازیابی اطلاعات فارسی است[31].

لیو و همکارانش فرآیند کلی برای استخراج کلمات کلیدی را ارائه کردند که در این فرآیند ابتدا کلمات کلیدی نامزد، تشخیص و به هر کلمه وزنی اختصاص داده شده و درنهایت کلمات کلیدی با بیشترین وزن انتخاب می‌شدند[25].

فرانتزی و همکارانش تحلیل آماری و زبان‌شناختی را با یکدیگر ترکیب کردند. آن‌ها معتقد بودند که اطلاعات آماری بدون اطلاعات زبان‌شناختی کلمات غیرمفید و غیرکلیدی را هم در نظر می‌گیرد [14].

Freitas and Kaestner در سال ۲۰۰۵ به‌دنبال پژوهش‌های قبلی و برای رفع مشکلات استخراج کلمات غیر کلیدی، با توجه به مجموعه‌ای از سندهای آموزشی و کلمات کلیدی مشخص برای آن‌ها، فرایند استخراج کلمات کلیدی را به‌عنوان یک مدل طبقه‌بندی کردند و کلمات بر اساس مشخصه‌هایی که دارند به "کلمات کلیدی" و "غیرکلیدی" طبقه‌بندی شدند و آن‌گاه احتمالات طبقه‌بندی به‌صورت آماری از مجموعه آموزشی یاد گرفته می‌شوند [16]. برای مثال Zhang در سال ۲۰۰۶ از الگوریتم طبقه‌بندی C 5.0 استفاده کرده است [39].

چویی و همکاران در سال ۲۰۱۴ [11] با استفاده n-gram و wordnet متون را در مورد تروریسم طبقه‌بندی کردند و به دقت قابل قبولی دست یافتند.

تانگیل و همکاران در سال ۲۰۱۴ [35] از طبقه‌بندی متن برای شناسایی بدافزارهای سیستم عامل اندروید استفاده کردند. آن‌ها طبقه‌بندی متن را برروی کد منبع برنامه‌های تحت اندروید اعمال کردند تا مشخص کنند که کدام کدهای منبع متعلق به بدافزارها هستند.

کولاس و همکاران در سال ۲۰۱۴ [9] روشی برای طبقه‌بندی متون به‌صورت تک برچسب[1] در حالتی که تعداد اسناد برچسب خورده کم باشند را ارائه دادند. روش آن‌ها برگرفته از تکنیک‌های Topic Modelling بوده و مبتنی بر الگوریتم LDA کار می‌کرد.

۲-۲- سامانه‌های طبقه‌بندی متون و اهمیت آن

در ده سال اخیر مدیریت مبتنی بر محتوای متون (که تحت عنوان کلی بازیابی اطلاعات شناخته می‌شوند) به علت رشد سریع و در دسترس قرار گرفتن متون به شکل دیجیتالی، از اهمیتی دوچندان برخوردار شده است. طبقه‌بندی متون یا عمل برچسب‌گذاری موضوعی متون زبان طبیعی، بر مبنای یک مجموعه از پیش تعیین‌شده، یکی از این موارد است. هم‌اکنون طبقه‌بندی متون در بسیاری از زمینه‌ها، از شاخص‌گذاری متون بر مبنای یک لغت‌نامه کنترل‌شده[2] ، فیلترکردن متون، تولید خودکار فراداده، ابهام زدایی از کلمه[3] ، تولید کاتالوگ‌های سلسله‌مراتبی از منابع وبی[4] و به‌طور کلی در هر کاربردی که نیاز به سازمان‌دهی مستندات یا توزیع انتخابی و تطبیقی خاصی از مستندات مد نظر باشد، کاربرد دارد.

از آن‌جا که برای ساختن طبقه‌ها نیازی به دانش مهندسی یا افراد خبره آن زمینه ندارد، مزایای این رهیافت، دقت قابل مقایسه در مقابل دقت به‌دست‌آمده از فرد خبره و کاهش قابل توجه در هزینه انسانی است.

[1] single label
[2] Controlled Dictionary
[3] Word Sense Disambiguation
[4] Population Of Hierarchical Catalogues Of Web Resources

طبقه‌بندی متون به‌صورت دستی علاوه‌بر زمان‌بری و هزینه زیاد، معایب زیر را نیز با خود به‌همراه دارد [17].

۱- برای زمینه‌های تخصصی خاص، نیاز به دانش افراد خبره دارد (مانند بانک‌های پزشکی، بانک‌های حقوقی)

۲- از آن‌جا که برچسب‌گذاری دستی مبتنی بر دانش و تجربه فرد می‌باشد، بسیار خطاپذیر است.

۳- تصمیم دو فرد خبره در برچسب‌گذاری می‌تواند متفاوت و حتی ناسازگار باشد (سیستم ناسازگاری درونی دارد) [10].

امروزه بنابر آن‌چه گفته شد، طبقه‌بندی متون در تقاطع یادگیری ماشین و بازیابی اطلاعات مطرح است. هم‌چنین برخی از مشخصات این مسأله، با مسائلی چون استخراج اطلاعات و دانش از متون و داده‌کاوی متون[1] مشترک است [23]. با این حال مرز و تعریف دقیق آنها کماکان مورد بحث می‌باشد. داده‌کاوی متون با تحلیل مقدار زیادی از متون و یافتن کاربرد الگوها، سعی می‌کند تا به استخراج احتمالی اطلاعات (تنها با استفاده از اطلاعات احتمالی) بپردازد. امروزه، این مسأله به‌طور روز افزون مورد استفاده قرار گرفته شده است. بر طبق این دیدگاه، طبقه‌بندی متون یکی از نمودهای داده‌کاوی متون است. مفاهیم دسته‌بندی متون هم‌اکنون از یک ادبیات عمیقی برخوردار گشته است؛ اما این مفاهیم عمدتاً پراکنده بوده است. باید متذکر شد، گهگاه طبقه‌بندی خودکار متون در مقاله‌هایی مورد استفاده قرار می‌گیرد که با آن‌چه در این‌جا مورد نظر است به‌طورکامل متفاوت است. از یک طرف تقسیم خودکار متون به طبقه‌های از پیش تعریف‌شده، که در این‌جا مد نظر است و از طرف دیگر تشخیص خودکار مجموعه‌ای از طبقه‌ها یا تشخیص خودکار مجموعه‌ای از طبقه‌ها و گروه‌بندی متون بر مبنای آن‌که به‌طورمعمول خوشه‌بندی[2] متون نامیده می‌شود و یا هر کاری که برای قراردادن متون در گروه‌های خاصی باشد، قابل گسترش است[27].

مسأله طبقه‌بندی متون به کار آقای مارون برای طبقه‌بندی متون به‌صورت احتمالی باز می‌گردد [28]. از آن پس برای کاربردهای متنوعی استفاده شده است. باید توجه داشت، از آن‌جا که بعضی از این گروه‌ها با هم هم‌پوشانی دارند، مرزهای بین این گروه‌ها دقیق نیست و بعضی از این گروه‌ها، بعضی دیگر را پوشش می‌دهند. از کاربردهای طبقه‌بندی متون به سامانه‌های خودکار پاسخ به سؤالات [30]، فیلترکردن اطلاعات، تشخیص موضوعیت داده‌ها و نامه‌های الکترونیکی بی‌ارزش، تشخیص عنوان و دیگر زمینه‌های مرتبط می‌توان اشاره کرد [30]. از دیگر کاربردها مواردی همچون طبقه‌بندی گفتاری[3] که ترکیبی از طبقه‌بندی متون و تشخیص گفتار[4] است [32]، طبقه‌بندی متون چندرسانه‌ای[5] از طریق عنوان‌های متنی [29]،[32] تشخیص نویسنده برای متون ادبیاتی نامشخص یا مورد بحث [15]، تشخیص زبان برای متونی که زبان آن‌ها نامشخص است [8]، تشخیص خودکار جنس متن[6] [22] و رتبه‌بندی خودکار کیفیت نوشتار[7] [24] است.

۳-۲- اصطلاح‌نامه

اصطلاح‌نامه مجموعه‌ای شامل واژه‌ها، اصطلاحات و اطلاعات مربوط به یک حوزهٔ خاص از معرفت بشری است. این مجموعه، واژگان زبان نمایه‌ای کنترل شده‌ای است که طوری سازمان یافته تا روابط پیشین میان مفاهیم (اعم و اخص و ...) را روشن کند [1]. واحد تشکیل‌دهنده اصطلاح‌نامه، واژه‌هایی است که تبلور اطلاعات و دربرگیرندهٔ مسائل متن و مدرک مورد نظر است، که این‌ها را دراصطلاح واژه‌ها یا نشانه‌های کلیدی و یا کلیدواژه می‌گویند. استخراج کلیدواژه از داخل متون و منابع به بازیابی اطلاعات متن کمک شایانی می‌کند. اصطلاح‌نامه، برای شاخص‌گذاری و جستجو مورد استفاده قرار می‌گیرد. برای استفاده افرادی که شاخص‌گذاری می‌کنند، ممکن است اصطلاح‌نامه‌ای ساده تهیه شود، که فقط فهرست الفبایی از واژه‌ها و اصطلاحات را با برقرارکردن یک نوع رابطه (رابطه هم‌ارز، مترادفات و شبه‌مترادفات) نشان دهد؛ ولی همان‌گونه که در تعریف اصطلاح‌نامه آمده است، باید اصطلاح‌نامه به شکل یک شبکه منطقی و سلسله‌مراتبی تنظیم شود که روابط معنایی و پیشین (در واقع موجود) میان اصطلاحات در یک حوزه علمی را در ساختاری منظم و گویا و در عین حال مستند و پویا ارائه کند. درواقع، هرگاه پژوهش‌گری اصطلاح‌نامه را به‌دست می‌گیرد و اصطلاحی را می‌یابد، روابط قبل و بعد و جایگاه اصلی و فرعی و بالا و

[1] Text mining

[2] Clustering

[3] Speech Categorization

[4] Speech Recognition

[5] Multimedia Document Categorization

[6] Automated Identification Of Text Genre

[7] Automated Essay Grading

پایین آن را به‌سهولت می‌بیند، و شکی نیست کـه شناسـایی جایگاه اصطلاح در علوم نقش مهمی در طرح‌ریزی موضـوع و پژوهش دارد. به‌طورعمومی از اصطلاح‌نامه برای طبقه‌بنـدی، شـاخص‌گـذاری، ذخیـره و بازیـابی اطلاعـات در بانک‌هـای اطلاعاتی می‌توان استفاده کرد.

۲-۴- اصطلاح‌نامه

رفع ابهام از کلمه[1]، یافتن درست کلمات هم‌نویس در یک متن است. برای مثال"مرد" در معنای اسمی به انسان ذکور بالغ می‌تواند گفته شود و یا در معنای فعلی بن ماضی سوم شخص ساده از مصدر "مردن" باشد. بنابراین یکی از وظایف رفع ابهام انتخاب یکی از این دو حالت برای این کاربرد در جمله ماست. رفع ابهام برای بسیاری از کاربردها هم‌چون بازیابی زبان‌های طبیعی، و شاخص‌بندی مستندات با استفاده از ترجیح نقش کلمه بر خود کلمه در حوزه بازیابی اطلاعات مورد توجه است. هم‌چنین رفع ابهام ممکن است به‌صورت یکی از وظایف طبقه‌بندی متون دیده شود[13].

۲-۵- بازیابی اطلاعات و استخراج کلمات کلیدی

بازیابی اطلاعات به‌طور اصولی مرتبط با بازیابی مستندات و مدارک است. کار معمول در بازیابی اطلاعات این است که بسته به نیاز مطرح‌شده از سوی کاربر، مرتبط‌ترین متون و مستندات را از میان دیگر مستندات یک مجموعـه بیـرون بکشد. این یافتن دانش نیسـت؛ بلکـه تنهـا آن تعـدادی از کلمات را که به‌نظرش مرتبط‌تر به نیاز اطلاعاتی جستجوگر است، به او تحویل می‌دهد. در بازیابی سند، اطلاعات، یک زیرمجموعـه از اسـناد هسـتند کـه در ظـاهر مـرتبط بـا پرس‌وجو است. تمام روش‌های جست‌وجو مبتنی بر مقایسه بین پرس‌وجو و سند ذخیره‌شده می‌باشند. بعضی مواقع این مقایسه به‌صورت غیرمستقیم و بـا مقایسـه پـرس‌وجـو بـا کلمات کلیدی انجام می‌گیرد[6].

۲-۶- شاخص‌گذاری خودکار

بر اساس استاندارد شاخص‌گذاری [26] (BS3700:1988)[2]شاخص‌ها مجموعه‌ای منظم از کلمات نشانه‌گذاری شده هستند تا کاربران را قادر سازند اطلاعاتی که محل آن‌ها در مدرک مشخص شده پیدا کنند. یکی از پیشگامان شاخص‌گذاری خودکار که طرحی بر پایه تشخیص کثرت حضور اصطلاح در مدرک ارائه کرده است، اچ. پی. لون است[26].

این طرح به‌طور کلی از سه مرحله تشکیل شده است:

۱- تعداد واقعی هر مفهوم در هر مدرک تشخیص داده می‌شود.

۲- بر اساس نتایج مرحله نخست تعداد هر مفهوم در مجموعه مدارک محاسبه می‌شود.

۳- در پایان همه مفاهیمی که در بیشینه تعداد ($Hmax$) یا در کمینه تعداد ($Hmin$) ظاهر می‌شوند، جدا و سپس لغاتی که باقی می‌مانند به‌عنوان شاخص استفاده می‌شوند (ارزش‌های $Hmax$ و $Hmin$ با روش‌های مناسب تعیین می‌شوند).

این مراحل در روش لون در نگاه نخست بسیار ساده و قابل درک به‌نظر می‌رسد. در این روش همه مفاهیمی که به تعداد زیادی در یک پایگاه اطلاعاتی وجود دارند، به‌عنوان شاخص انتخاب نمی‌شوند. هم‌چنین در این روش همه مفاهیمی که به‌ندرت در مدرک وجود دارند، بدون بار معنایی تلقی و فرض می‌شود که توان ارائه محتوای مدرک را ندارند. مزیت این روش در واقع‌گرایی آن و ضعف آن در تعیین فقط بیشترین و کمترین تعداد لغات است. به‌عبارت دیگر فرض بر آن است که مفاهیمی که اغلب در مجموعه داده‌ها وجود دارند، شاخص نیستند، چون با این مفاهیم مدارک را از یکدیگر نمی‌توان تفکیک کرد و مفاهیمی که به‌ندرت در مدارک می‌آیند، قابلیت ارائه محتوای مدارک را ندارند. با ضوابط مطرح‌شده، طرح لون قواعد زیر را به‌دنبال دارد:

در این طرح از مرحله نخست دو عملکرد زیر امکان پذیر است:

۱- پس از مقایسه مدارک با اصطلاحات یک فهرست کلمات توقف، همه لغاتی که در این فهرست قرار دارند، انتخاب و حذف می‌شوند.

۲- مفاهیم باقی‌مانده بر اساس ریشه لغت تفکیک می‌شوند. بنابراین مفاهیم خانه، خانه‌ها بر اساس شکل خانه تفکیک می‌شوند.

در فهرست کلمات توقف یادشده حروف ربط، حروف تعریف، لغات دیگری از این قبیل وجود دارند که در متن بسیار زیاد هستند. آزمایش‌ها نشان می‌دهند که ۲۰٪ از بیشترین تعداد لغات متن به‌طورتقریبی ۷۰٪ از متن را شامل می‌شوند. مزیت روش لون واقع‌گرایی است؛ زیرا وظیفه اصلی

[1] Word Sense Disambiguation (WSD

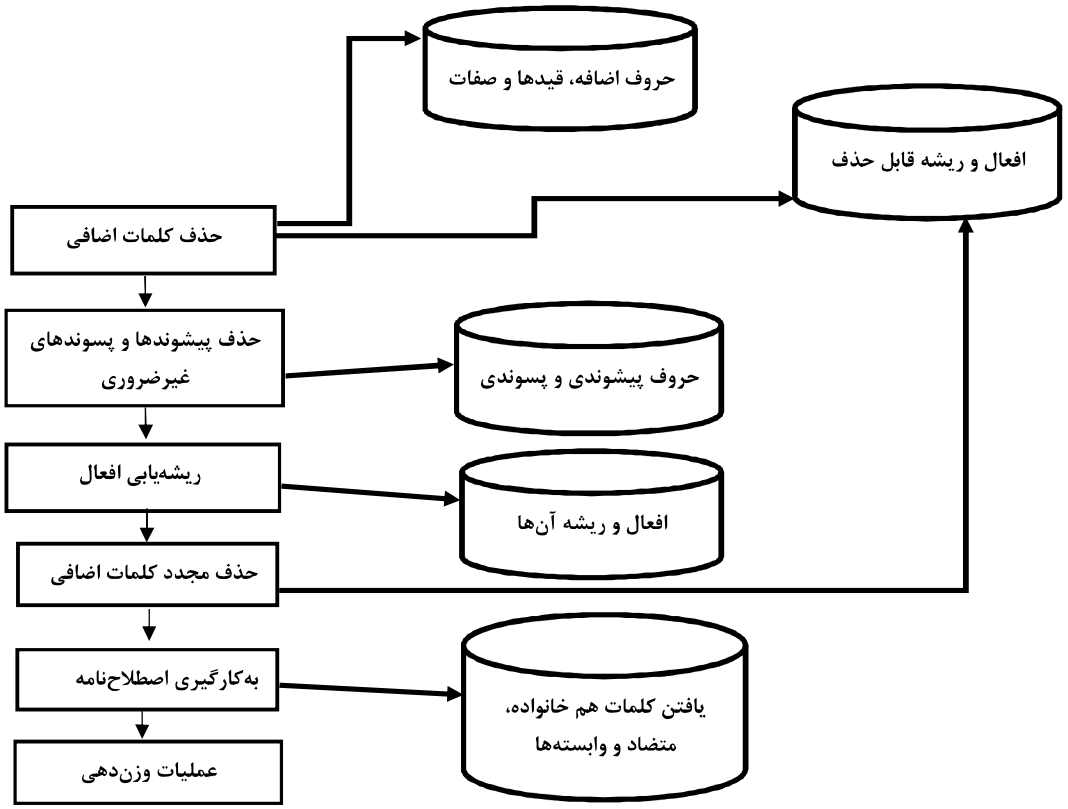
[2] British Indexing Standard

محاسبه تعداد اصطلاحات و تفکیک آن‌ها بر اساس تعدادشان و مشکل این روش در تعیین مرز $Hmin$ و $Hmax$ است. حذف مفاهیمی با بیشینه تعداد جامعیت را می‌تواند پایین بیاورد و باعث شود که بسیاری از مدارک مرتبط نادیده گرفته شوند. حذف مفاهیمی با کمینه تعداد مانعیت را می‌تواند پایین بیاورد. سالتون بیان می‌کند که تا چه حد یک مفهوم در یک مدرک از یک مدرک دیگر قابل تشخیص است. در این روش در ابتدا عامل شباهت برای دو مدرکی محاسبه می‌شود که شباهت میان آن‌ها باید مشخص شود. این شباهت هنگامی به عدد یک می‌رسد که هر دو مدرک به‌طورکامل مطابق یکدیگر شاخص‌گذاری شده‌اند. عامل شباهت برای هر جفت مدرک در نظر گرفته می‌شود و در پایان یک شباهت نسبی میان کل مدارک و انبوه مدارکی که مورد بررسی قرار می‌گیرند، حاصل و تفکیک آماری همه مفاهیم با یک کثرت نسبی محاسبه می‌شود. تفکیک آماری به‌صورت مثبت: برای مفاهیمی با بسامد متوسط باید هنگام شاخص‌گذاری ارجحیت قائل شویم. تفکیک آماری به‌صورت منفی: مفاهیمی که با تعداد زیاد در مدارک وجود دارند و باعث شباهت مدارک به یکدیگر می‌شوند. چنین بسامد بالایی برای شاخص‌گذاری به‌هیچ‌وجه مناسب نیست [33].

براساس فرضیه هارتر در دو سطح با دو احتمال متفاوت برای یک اصطلاح تقسیم می‌شوند؛ سطح نخست برای مدارکی است که در آن‌ها یک اصطلاح وجود ندارد و سطح دوم برای مدارکی است که یک اصطلاح در آن حضور قابل توجهی دارد. هارتر نشان داد که احتمالات در سطوح نخست اصطلاحات ناخواسته و بدون بار معنایی و در سطح دوم اصطلاحات معنی‌دار را بازیابی می‌کنند. بنابراین باید از روش‌های آماری مناسب و صحیح استفاده شود [19].

۳- راه‌کارهای پیشنهادی

در این بخش، روش پیشنهادی به‌منظور خودکارسازی طبقه‌بندی متون فارسی به‌کمک اصطلاح‌نامه ارائه شده است. نمودار فرایند پیشنهادی در شکل (۱) نشان داده می‌شود. در ادامه این بخش تک‌تک مراحل معرفی‌شده در نمودار به تفصیل مورد بررسی قرار می‌گیرد.



(شکل-۱): روش پیشنهادی
(Figure-1): Proposed Method

۱-۳- تحلیل واژگانی

متون الکترونیکی فارسی موجود به‌صورت خام دارای مشکلات زیر هستند:

۱- عدم وجود رسم‌الخط یکسان: متأسفانه متون فارسی موجود از یک رسم‌الخط واحد پیروی نمی‌کنند. یکی از مهم‌ترین این دلایل، متداول‌بودن بعضی از حالت‌های چسبیدن وندهایی همانند" ها "به اسم‌ها در زبان فارسی است. برای مثال "کتابها"،" کتاب ها"، "کتاب‌ها. "

۲- وجود کدهای متفاوت برای حروف فارسی: اگرچه سیستم یونی‌کد، سعی در یکسان‌سازی کدها برای کلیه زبان‌ها کرده است، ولی با این حال در مواردی برای پشتیبانی از کلیه حالت‌ها ناچار به استفاده از کدهای بعضا متفاوت برای یک حرف کرده است.

۲-۳- حذف کلمات اضافی

واژه‌های عمومی زبان در شاخص‌گذاری ارزش کمی دارند؛ به‌همین علت روند کلی شاخص‌گذاری با حذف واژه‌های عمومی آغاز می‌شود. از دلایلی که سبب می‌شود حروف اضافی را حذف کنیم این است که این‌ها کمکی به طبقه‌بندی نمی‌کنند و این‌که حجم بالایی دارند. این واژه‌ها حدود سی تا پنجاه درصد متون را تشکیل می‌دهند. درواقع این واژه‌ها، واژه‌هایی هستند که کاربر مایل به جستجوی آن‌ها نیست. حذف کلمات اضافی نظیر حروف اضافه، بسیاری از قیدها و صفات، برخی از افعال (که خود ریشه‌اند)، حروف ربط و غیره به‌طورعمومی در مضمون کلی متن تأثیر منفی نمی‌گذارد، بلکه باعث خلاصه‌سازی متن می‌شود؛ به‌همین دلیل این کلمات را در یک فایل متنی (شکل ۲) می‌توان وارد کرده، در هنگام اجرای برنامه این کلمات از فایل خوانده شده و در ساختمان سریع جدول درهم‌سازی، ریخته می‌شوند (این ساختار نیاز به جستجوی ترتیبی ندارد). هنگام بررسی هر مستند فارسی، کلمات اضافی موجود در این صفحه با جستجو در جدول درهم سازی ، شناخته شده و از مستند حذف می شوند. نکته قابل توجه این است که در این مرحله، ما تنها افعال ریشه را از مستند حذف می‌کنیم. حذف سایر افعال که ریشه نیستند، پس از مرحلهٔ ریشه‌یابی انجام می‌شود.

۳-۳- ریشه‌یابی کلمات فارسی

برای ریشه‌یابی کلمات انگلیسی از الگوریتم‌های مختلفی استفاده می‌شود که مشهورترین آن‌ها الگوریتم Porter نام

است
هست
.....
من
ما
...
از
و
...

(شکل-۲): فایل متنی شامل کلمات اضافی که حذف آن‌ها در مضمون کلی متن اثر ندارد.

(Figure-2): The text file contains stop word that have no effect on the overall text content.

دارد. کلمات انگلیسی دارای پیشوندها و پسوندهای فرموله‌شده‌ای هستند که با استفاده از این الگوریتم‌ها، امکان جداسازی آن‌ها از ریشه کلمه وجود دارد. برای مثال:

فعل + ing → استمراری فعل (go + ing →going)

ریشه + ness → اسم (soft + nesss →softness)

با وجود ساخت‌یافتگی نسبی کلمات انگلیسی، این زبان استثناهایی هم دارد که الگوریتم Porter قادر به شناسایی آن‌ها نیست برای مثال کلمه "Bus" در این الگوریتم به‌اشتباه یک کلمه جمع شناخته شده و "s" آن حذف می‌شود که خود ریشه است.

Bus-> Bu + s

اما دستور زبان فارسی استثناهای بیشتری دارد، که با استفاده از الگوریتم‌های فرموله‌بندی‌شده قابل تشخیص نبوده و در نتیجه امکان ریشه‌یابی آن‌ها به‌طور صحیح وجود ندارد. قبل از آن‌که بیشتر به الگوریتم ریشه‌یابی پیشنهادی پرداخته شود، باید مشکل مرز بین واژگان حل شود. امروزه تمام سامانه‌های کامپیوتری ذخیره و بازیابی اطلاعات، با زبان فارسی از نظر فاصله میان واژه‌ها و عبارت‌ها مشکل دارند. درصورتی‌که خط فارسی صدها سال است که توانسته ارتباط میان افراد را برقرار کند و هیچ مشکل غیرقابل حلی هم نداشته باشد. مشکل از جایی شروع می‌شود که پای فناوری نوین به میان می‌آید. بنابراین می‌توان نتیجه گرفت که راه‌حل این مشکل را یا باید در رسم الخط فارسی و یا در فناوری نوین یافت.

یکی از منطقی‌ترین راه‌حل‌ها برای رفع این مشکل آن است که منتظر باشیم فرهنگستان زبان و ادب فارسی رسم‌الخط دقیق و مدون فارسی را ارائه کند. اگر چنین رسم‌الخطی تدوین شود، به‌حتم مشکل فاصله میان واژه‌ها و عبارت‌ها نیز حل خواهد شد. در آن زمان به‌احتمال همه یکسان عمل خواهند کرد و مشکلات رسم‌الخط فارسی

کاهش خواهد یافت. ویژگی عمده این راه‌حل آن است که هم با سامانه‌ها و کاربران از یک رسم‌الخط تبعیت می‌کنند و میزان ناهماهنگی در پایین‌ترین حد خواهد بود. این راه‌حل یک مشکل اساسی دارد، یعنی معلوم نیست چه زمان رسم‌الخط یکسان‌شدهٔ نهایی تدوین خواهد شد. در مورد پیوسته‌نویسی یا جدانویسی ده‌ها سال است که بحث و گفتگو وجود دارد و هنوز به نتیجه قطعی نرسیده است. حتی اگر در زمانی خاص نتیجه نهایی به‌دست آید و رسم‌الخط دقیق فارسی مشخص شود، برای این که همه آن را اجرا کنند باز هم به سالیان بسیاری وقت نیاز است.

راه‌حل دیگری نیز برای حل مشکل رسم‌الخط فارسی پیشنهاد شده است. در این راه‌حل نرم افزارها به‌گونه‌ای طراحی می‌شوند که هرگاه واردکننده اطلاعات به بود یا نبود فضای خالی میان واژه‌ها شک داشته باشد با زدن علامتی میان آن واژه‌ها، ورود اطلاعات را ادامه می‌دهد. اما به‌نظر نمی‌رسد، چنین پیشنهادی بتواند مشکل مرز بین واژه‌ها را حل کند. با تمام این احوال، به‌کارگیری پیشنهادهای زیر معقول به نظر می‌رسد:

۱- قبل از هر واژه فاصله خالی درج شود.

۲- بین حروف اصطلاحات و واژه‌های لاتین که به‌طوردقیق منعکس‌کنندهٔ لفظ خارجی است، فاصله خالی درج نشود.

۳- انواع پیشوندها و پسوندها، پیوسته یا بدون فاصله خالی نوشته شوند. اگر نیاز باشد تا پسوند یا پیشوند به واژه اصلی به‌طورکامل متصل نشود (به‌عنوان مثال پسوند **اند** هنگام اتصال به واژه **شده**، به‌جای **شده‌اند** به‌صورت **شده‌اند** درآید). باید از نویسه‌های کنترلی موجود در نرم‌افزارهای حروف‌چینی استفاده و از درج فاصله خالی اجتناب کرد.

می‌دانیم که هر قانونی که برای ریشه‌یابی به‌کار رود، به‌راحتی بی‌قاعدگی‌ها و استثنائات فراوانی برای آن قانون می‌توان یافت. برای حل چنین مشکلی دو راه حل به‌نظر می‌رسد. یکی این‌که فهرستی از استثنائات نگهداری شود؛ دیگری این‌که از یک فهرست برای تشخیص واژه‌های درست استفاده شود. با توجه به فراوانی استثنائات در قوانین ریشه‌یابی، راه‌حل دوم در عمل مناسب‌تر به‌نظر می‌رسد. اکنون به الگوریتم پیشنهادی، برای ریشه‌یابی پرداخته می‌شود. در این الگوریتم ابتدا واژه در فهرست واژگان که فقط اسامی جامد را در بر می‌گیرد و از قبل آماده شده است، جستجو می‌شود. در صورت وجود واژه در فهرست واژگان عملیات ریشه‌یابی انجام نمی‌شود و خود واژه به‌عنوان ریشه اصلی در نظر گرفته می‌شود؛ اما درصورتی‌که واژه در فهرست واژگان موجود نباشد، تا حد ممکن الگوریم ریشه‌یابی روی آن واژه اجرا می‌شود. چنان‌چه واژه حاصل‌شده در هر مرحله از الگوریتم، در فهرست واژگان موجود بود، عملیات مورد قبول واقع می‌شود و کار به پایان می‌رسد. در طراحی ریشه‌یاب سعی بر آن بوده است که ضمن حفظ انسجام، اجزاء مستقل از یکدیگر باشند، بنابراین سامانه از چندین بخش تشکیل شده است، این بخش‌ها عبارتند از:

الف) مجموعه قوانین ریشه‌یابی (برای زدودن پسوند و پیشوند) که با استفاده از قوانین ساخت واژه در زبان فارسی استخراج شده‌اند؛

ب) فهرست مصادر فارسی (اعم از افعال ساده، پیشوندی، و مرکب و نیز عبارت های فعلی)؛

ج) فهرست واژگان جمع مکسر که برای هر واژه جمع مکسر، مفرد آن را مشخص می‌کند.

د) فهرست واژگان جامد که کلیه واژگان فارسی به غیر از افعال و مشتقات آن، جمع‌های مکسر و نیز مابقی واژگان مشتق را نگهداری می‌کند. به‌دلیل این‌که چنین فهرستی برای زبان آزمایشی را، که اکنون حدود نه هزار واژه را در بردارد، می‌توان به مرور تکمیل‌تر کرد. برای آن‌که بتوان ریشه واژگانی را که از زبان عربی وارد فارسی شده‌اند، استخراج کرد، دو روش به‌نظر می‌رسد. یکی این‌که قواعد و ریشه‌های واژگان زبان عربی را نیز، هر چند محدود، در سامانه وارد و یا این‌که با استفاده از یک فرهنگ ترادف تمامی واژگان هم‌خانواده را فهرست کنیم.

۴-۳- حذف مجدد کلمات اضافی

در این مرحله به‌طورتقریبی تمام افعال و اسامی با ریشه‌ها جایگزین شده‌اند. حال طی یک مرحله، دوباره کلمات اضافی، بر اساس فایل کلمات اضافی که در مرحله اول استفاده شد، حذف و برخی از افعال پس از حذف شناسه به افعال ریشه تبدیل می‌شوند که قابل حذف است. در پایان این مرحله بخش عظیمی از کلمات اضافی از متن حذف می‌شوند. فهرست واژگانی که در جدول (۱) آمده است توسط دکتر بی‌جن‌خان ایجاد شده است که حدود هشتصد لغت است که تنها بخشی از آن در جدول (۲) آمده است.

(جدول-۲): فهرست واژه‌های عمومی(stop word)

(Table-2): Stop word list

آن	ای	به	چندگانه	خودشان	روی	گرچه	میشوند	هر	همیشه
آنان	ایشان	بیشتر	چندین	خودم	زیرا	گرفته	میکند	هریک	همین
آنجا	این	بین	چنین	خودمان	سپس	لکن	میکنند	هست	هنوز
آنچه	اینجا	پس	چه	خویش	شامل	لیکن	نظر	هستند	هیچ
آنکه	اینست	تا	چون	داده	شاید	ما	نمیتوان	هستیم	هیچکدام
آنگاه	اینگونه	تایی	چیز	دارای	شد	مابین	نوع	هم	هیچگونه
آنها	اینها	تو	چیزی	دارد	شده	مانند	نیاز	همان	و
از	با	توسط	چیست	داشته	شما	مختلف	نیز	همانطور	وجود
است	باشد	چرا	حتی	در	شود	من	نیست	همانند	وگرنه
اگر	باید	چگونگی	خواهد	درباره	صورت	مورد	نیستند	همچنین	ولی
اگرچه	بدون	چگونه	خواهیم	دو	فقط	میباشد	ها	همچون	وی
الان	بر	چنان	خود	دیگر	کدام	میتوان	های	همدیگر	یا
اما	برای	چنانچه	خودت	دیگران	کرد	میتواند	هایی	همه	یک
انجام	بنابر	چند	خودتان	دیگری	کردن	میدهد	هر	همواره	یکدیگر

۵-۳- به‌کارگیری اصطلاح‌نامه

در این مرحله با استفاده از اصطلاح‌نامه برای ویژگی‌های اصلی استخراج‌شده از بخش پیش‌پردازش، کلمات هم‌خانواده، مترادف، و وابسته‌ها (اعم و اخص) شناسایی می‌شوند. به‌عبارت دیگر، برای تک‌تک کلمات اصلی متن، کلمات هم‌خانواده، مترادف، متضاد، اعم و اخص و وابسته استخراج و در جایی نگهداری می‌شود. از این کلمات بعدها برای وزن‌دهی استفاده خواهد شد. هدف آن است که درصورت دیدن کلمات هم‌خانواده در متن، به‌جای آن‌که به‌صورت مجزا برای هر یک وزنی در نظر گرفته شود، یک کلمه از میان آن‌ها به‌عنوان نماینده انتخاب شده و مجموع وزن کلمات هم‌خانواده و وابسته به نماینده در متن براساس یک ضریب وزنی مشخص به وزن کلمه نماینده اضافه می‌شود.

روش استخراج کلمات کلیدی از متن از یک زاویه به دو صورت است:

۱- شاخص‌گذاری مبتنی بر اصطلاح‌نامه

۲- شاخص‌گذاری مبتنی بر متن کامل

شاخص‌گذاری مبتنی برمتن کامل، بر اساس کلماتی است که در متن خود مدرک آمده است؛ ولی در شاخص‌گذاری مبتنی بر اصطلاح‌نامه، ممکن است واژه‌ای برای شاخص‌گذاری انتخاب شود که در متن نیامده باشد؛ ولی هم‌خانواده‌های آن‌ها موجود باشد. در این مقاله کلمات انتخاب‌شده براساس اصطلاح‌نامه می‌آید.

۱-۵-۳- رابطه مترادف

در کارهای گذشته درخصوص کلمات و عبارات هم‌معنی و مترادف پژوهش‌هایی صورت گرفته و بر اساس نتایج آن برای کلمات مترادف ضریب وزنی "۱" محاسبه شده که به‌طوردقیق معادل وزن خود کلمه است. یعنی بیشینه وزنی که به یک کلمه می‌توان اختصاص داد.

۲-۵-۳- رابطهٔ هم‌خانواده

اگر یک متن را در نظر بگیرید، اغلب اوقات، از کلمات هم‌خانواده استفاده می‌شود. مثل متنی که حاوی کلمات دلیل، دلالت و استدلال است. در واقع این کلمات همگی هم‌خانواده یک کلمه است. وجود کلمات هم خانواده باعث پراکندگی می‌شود. اشکال اصلی روش‌های موجود در بازیابی متن در زبان فارسی، عدم توجه به کلمات هم‌خانواده است. اگر خودمان در متن مذکور بخواهیم کلمات کلیدی را وزن دهیم با خواندن متن مذکور وزن بالایی به کلمه دلیل می‌دهیم. ما می‌خواهیم عملکرد مغزی را خودکار کنیم. پس این روش قادر است با دیدن کلمات هم‌خانواده، همه را به‌عنوان کلمه اصلی در نظر گرفته و بیشترین وزن را به کلمه اصلی دهد و بدین ترتیب روش پیشنهادی از توزیع وزن کلمات هم‌خانواده که مانع از تشخیص دقیق طبقه متن مورد نظر می‌شود جلوگیری می‌کند. برای انجام این‌کار با استفاده از اصطلاح‌نامه برای کلمات اصلی موجود در متن، کلمات هم‌خانواده پیدا شده و با دیدن هر یک از کلمات در متن، وزن کلمه اصلی به میزان یک ضریب وزنی مشخص اضافه می‌شود؛ یعنی یک کلمه به‌عنوان نماینده تعیین و هر بار به وزن این کلمه اضافه می‌شود که هدف در این پژوهش یافتن یک ضریب وزنی با دقت بالا برای کلمات هم‌خانواده است.

مثال: در شکل (۳) متنی حاوی واژه‌های هم‌خانواده عنوان شده است.

در محیط پیرامون ما هر روز حوادث زیادی اتفاق می‌افتد. بروز هر حادثه الزاما دلایلی دارد که باید در مورد هر دلیل تحقیق و بررسی لازم صورت گیرد. ممکن است افراد مختلف استدلال‌های متفاوتی ارائه نمایند. لزوما دلیل‌ها همیشه واضح و قابل مشاهده نخواهند بود و نیاز به بررسی بیشتر دارد.

(شکل-۳): متن حاوی واژه‌های هم‌خانواده
(Figure-3): Text containing family terms

در متن بالا داریم :
(دلایل : ۱ بار، استدلال : ۱ بار، دلیل : ۱ بار)

اگر فقط بخواهیم تعداد تکرارهای کلمات موجود در متن را در نظر داشته باشیم، کلمات کلیدی مناسبی را انتخاب نخواهیم کرد. درحالی‌که با رجوع به اصطلاحنامه و درنظرگرفتن کلمات هم‌خانواده و وابسته، به‌طور مثال با دیدن کلمه "دلیل"، مشخص می‌شود که این کلمه هم خانواده "دلایل" است و لذا به کلمه مرجح آن یعنی "دلایل" یک ضریب وزنی مشخص اضافه می‌شود. هم‌چنین در مورد کلمه "استدلال"، باعث می‌شود یک ضریب وزنی مشخص به فرکانس کلمه "دلایل" اضافه شود. یعنی درنهایت وزن "دلایل" در این متن به‌میزان دو ضریب وزنی اضافه می‌شود.

۳-۵-۳- رابطه اعم

اگرچه استفاده از کلمات هم‌خانواده با استفاده از اصطلاحنامه در وزن‌دهی کلمات موجود در متن کمک زیادی به فرآیند بازیابی و طبقه‌بندی متن می‌کند، اما لزوماً برای طبقه‌بندی صحیح متن‌ها در متون کافی نیست. در شکل (۴) مثالی در مورد آموزش و پرورش ارائه شده است:

نظام جدید آموزش متوسطه از سال ۱۳۷۱ در کشور آغاز گردید. وزارت آموزش‌وپرورش در جهت تاثیرگذاری مثبت بر سطح علمی و آموزش دانش‌آموزان کشور اقدامات زیادی انجام داد. در ابتدا برخی از مدارس کشور ملزم اجرای این طرح شدند. در ادامه در خصوص مدارس راهنمایی و دبستان‌ها نیز نظام آموزشی جدیدی وضع گردید که دروس جدیدی جهت تدریس در پایه‌های مختلف تحصیلی اضافه گردید.

(شکل-۴): متن حاوی واژه‌های اعم
(Figure-4): The text contains general terms

در این متن کلمات مدارس، آموزشگاه، دبستان، آمـوزش و..... بیـــش از سایر کلمات تکرار شده‌اند؛ اما همان‌طورکه ملاحظه می‌شود، این کلمات هم‌معنی نیستند و یک کلمه را به‌عنوان نماینده همه آن‌ها نمی‌توان در نظر گرفت؛ اما در اصطلاحنامه مورد استفاده همه این کلمات در زیر رده آموزش و پرورش قرار می‌گیرد. درصورت عدم استفاده از اصطلاحنامه ممکن است، طبقه‌بندی صحیح متن مورد نظر به‌دلیل پراکندگی کلمات اصلی امکان‌پذیر نباشد؛ اما در روش پیشنهادی برای هر یک از این کلمات، اعم آنها که همان آموزش است در نظر گرفته شده و به این‌ترتیب کلمه کلیدی آموزش امتیاز بالایی در متن پیدا می‌کند.

۴-۵-۳- رابطه اخص

در این حالت از کلمات خیلی کلی‌تر به جزئی‌تر می‌رسیم. مثال شکل (۵) در مورد جانوران است.

جانوران به‌طور کلی به دو گروه مهره‌داران و بی‌مهرگان تقسیم می‌شوند. مهره‌داران نیز خود به ۵ گروه مجزا شامل ماهی‌ها، دوزیستان، خزندگان، پرندگان و پستانداران تقسیم می‌شوند که هر یک از این گروه‌ها به‌نوبه خود زیر گروه‌هایی دارند. به‌طور مثال سفره‌ماهی در زیرگروه ماهی‌های غضروفی است و ماهی سفید در زیرگروه ماهی‌های استخوانی قرارمی‌گیرد.

(شکل-۵): متن حاوی واژه‌های اخص
(Figure-5): The text containing specific terms

(جانوران: ۱ بار، مهره‌داران: ۱ بار، ماهی‌ها: ۵ بار)

در متن بالا که در آن کلمات جانوران، ماهی‌ها، پستانداران و... استفاده شده است، این‌ها کلمات کلی هستند. از این کلمات متوجه می‌شویم که متن راجع به جانوران است؛ ولی وارد زیر شاخه می‌شویم تا متوجه شویم به‌صورت جزئی به کدام دسته‌بندی اشاره دارد. اگر به اخص آن‌ها اشاره کنیم به کلمه ماهی‌ها برمی‌خوریم و چون ماهی‌ها با وزن معقولی در متن هم وجود داشت، می‌فهمیم به‌طور کلی متن در مورد کلمه ماهی است. بدین ترتیب می‌توانیم متن‌ها را با دقت بالاتری طبقه‌بندی کنیم و با همان دقت بالا به پرس‌وجوهای کاربر پاسخ صحیح دهیم. پس اگر کاربری به‌طور خاص راجع به متون ماهی‌ها درخواست می‌دهد، در درجه نخست متن‌های مستقیم با ماهی در اختیار او قرار خواهد گرفت و درجه‌های بعدی متون مربوط به مهره‌داران و جانوران به آنها پاسخ داده خواهد شد.

هدف در این پژوهش هم‌چنین، ارائه یک الگوریتم معین جهت یافتن یک ضریب وزنی مشخص و دقیق برای کلمات دارای رابطه اعم و اخص است.

۶-۳- مرحله وزن‌دهی

۱-۶-۳- وزن‌دهی برای کلمات هم‌خانواده

برای این نوع کلمات وزن‌دهی به این‌ترتیب انجام می‌شود که برای کلمات هم‌خانواده موجود در متن به‌کمک اصطلاحنامه،

ضریب وزنی Beta در نظر گرفته می‌شود. اگر چند کلمه هم‌خانواده در متن وجود داشته باشند، براساس اصطلاح‌نامه به‌ازای یافتن هر واژه هم‌خانواده برای یک لغت به میزان درصد وزنی Beta به وزن لغت اضافه می‌شود.

۲-۶-۳- وزن‌دهی کلمات اعم و اخص

با توجه به این‌که برای کلمه اعم و اخص یک کلمه موجود در متن وزن مساوی با آن را نمی‌توان اختصاص داد؛ لازم است، روش ساخت‌یافته‌ای برای وزن‌دهی کلمات موجود در متن و وابسته‌های آن‌ها ارائه شود. از آن‌جایی‌که اعم یک کلمه اصلی در متن لزوماً نمی‌تواند جای‌گزین آن کلمه شود و درصورت جای‌گزینی با کلمه اصلی مفهوم اصلی متن می‌تواند تغییر کند، نمی‌توان وزن مساوی برای کلمات اعم و اخص و کلمه نماینده در نظر گرفت. در چنین مواردی بر اساس نوع کاربرد و میزان دقت مورد نیاز در طبقه‌بندی وزنی کمتر یک برای کلمات اعم یک کلمه اصلی در نظر گرفته می‌شود. برای مثال برای هر سطح در ساختار درختی وزن‌های به‌ترتیب ضریبی از Alpha در نظر گرفته می‌شود که میزان Alpha در ادامه مطرح می‌شود.

۳-۶-۳- طبقه‌بندی متون

طبقه‌بندی متون به‌معنای ایجاد تناظر بین یک سند و مجموعه‌ای از طبقه‌های ازپیش‌تعریف‌شده است. مراحل زیر را برای یک طبقه‌بندی‌کننده متون می‌توان در نظر گرفت:

- پردازش و بازنمایی اسناد: ایجاد یک بازنمایی عددی برای هر سند؛
- انتخاب ویژگی‌ها: این مرحله شامل انتخاب زیرمجموعه‌ای از کلمات موجود در مجموعه اسناد است که بهترین نحو توانایی بازنمایی این مجموعه را داشته باشند و هدف از آن بهبود کارایی وسرعت طبقه‌بندی‌کننده است؛
- وزن‌دهی به کلمات: نسبت‌دادن یک مقدار عددی مناسب به هر یک از ویژگی‌های انتخاب‌شده تا تمایز هر سند از سایر اسناد نمود بیشتری داشته باشد؛
- آموزش طبقه‌بندی‌کننده: آموزش طبقه‌بندی‌کننده با استفاده از بازنمایی عددی اسناد آموزشی حاصل از مرحله قبل.

طبقه‌بندی متون به‌طور عمیقی متکی بر اصول حاکم بر بازیابی اطلاعات است. طبقه‌بندی متون یک عمل مدیریت متون مبتنی بر مفهوم است. به‌طوری‌که مشخصه‌های مشترک بسیاری با اعمال بازیابی اطلاعات، هم‌چون جستجوی متن دارد. در این مرحله، پس از مشخص‌شدن وزن کلمات موجود در هر متن، نسبت به یافتن شاخص کلیدی متن اقدام می‌شود. برای این منظور کلماتی که دارای وزن بیشتری باشند، بیان‌گر این مفهوم هستند که اهمیت بیشتری در متن دارند و به‌عنوان شاخص کلیدی می‌توانند انتخاب شوند. پس از مشخص‌شدن شاخص‌های کلیدی و مهم در متون با توجه به نوع کاربرد موضوعی این شاخص‌ها، متون در دسته‌های ازپیش‌تعیین‌شده قرار می‌گیرند.

۴- ارزیابی روش پیشنهادی

به‌منظور پیاده‌سازی روش پیشنهادی از روش طبقه‌بندی استفاده شده است. به این‌منظور، یک فضای برداری شامل تمامی کلمات موجود در پایگاه داده‌ای متن ایجاد می‌شود؛ درنتیجه هر متن به‌صورت برداری از کلمات در فضای برداری نمایش داده می‌شود که هر متن نقطه‌ای از این فضای برداری خواهد بود. با استفاده از روش طبقه‌بندی، متون مشابه و مربوط به یک موضوع، در یک دسته قرار می‌گیرد. پرس‌وجوی کاربر نیز به‌سادگی با مقایسه با مراکز دسته‌ها پردازش‌شده و نزدیک‌ترین دسته شامل شامل شبیه‌ترین متن‌های موجود در طبقه به کاربر برگردانده می‌شود. در این روش از الگوریتم نزدیکترین همسایه یا KNN برای طبقه‌بندی استفاده می‌شود. الگوریتم KNN یک الگوریتم تعلیم با سرپرستی است. در حالت کلی از این الگوریتم به دو منظور استفاده می‌شود: یکی برای تخمین تابع چگالی توزیع داده‌های آموزش و دوم برای طبقه‌بندی داده‌های آزمایش بر اساس الگوهای آموزش داده‌شده. برای تخمین $p(x)$ از روی n نمونه تعلیم توسط الگوریتم KNN یک سلول به مرکزیت x می‌توانیم ایجاد کرده و اجازه دهیم، شعاع این سلول تا حدی گسترش پیدا کند که kn نمونه تعلیم را در بر گیرد. این نمونه‌ها kn نزدیکترین همسایه‌های x هستند. در حالت کلی k را به‌صورت kn در نظر می‌گیریم که kn تابعی تعریف‌شده از n است. اگر چگالی نقاط تعلیم اطراف x زیاد باشد، سلول کوچک می‌شود و بنابراین نتیجه به‌دست‌آمده نتیجه بهتری است و درصورتی‌که چگالی نقاط تعلیم اطراف x کم باشد، سلول بزرگ می‌شود. روش KNN، x را در دسته‌ای طبقه‌بندی می‌کند که بیشترین تکرار را در بین k نزدیکترین همسایه x دارد؛ درضمن برای یافتن میزان ارتباط بین

کلمات از ایده‌ای که در مقاله‌ای در سال ۲۰۱۰ توسط آقای George Tsatsaronis و هم‌کاران‌شان ارائه شده است، استفاده می‌شود[36].

۱-۴- نحوه انتخاب متون فارسی جهت آزمایش

به‌منظور آزمایش روش پیشنهادی، مجموعه‌مقالات روزنامه هم‌شهری در پنج دسته مختلف از وب‌گاه هم‌شهری[1] گردآوری شده است. اطلاعات کلی این مقالات در جدول (۳) ارائه شده است.

(جدول-۳): اطلاعات اولیه متون

(Table-3): Basic Information of Texts

ردیف	دسته بندی موضوعی	تعداد مقالات	متوسط تعداد کلمات مقالات
۱	ورزشی	۱۴۶	۲۰۴
۲	اقتصادی	۱۵۴	۱۹۹
۳	شهری	۱۷۱	۱۲۳
۴	حوادث	۸۹	۱۶۰
۵	خارجی	۱۳۰	۱۷۷

مجموعه‌مقالات به‌گونه‌ای انتخاب شده‌اند که دایره وسیعی از کلمات مربوط به هر حوزه را پوشش دهند. به‌عبارت بهتر، از هر دسته، مقالاتی با نویسندگان مختلف انتخاب شده که از دایره لغت‌های مختلفی برای نوشتن مقالات استفاده می‌کنند. به این‌ترتیب، توانایی راه‌کار پیشنهادی در شناسایی متون مختلف با کلمات گوناگون، اما در یک حوزه مشخص بهتر نشان داده می‌شود.

۲-۴- مراحل اجرا

همان‌گونه که در فصل قبل توضیح داده شد، این فرایند شامل چند مرحله اصلی است که در ادامه هر یک از مراحل شرح داده می‌شود.

۱-۲-۴-مرحلهٔ پیش پردازش

همان‌گونه که بیان شد، در این مرحله بایستی کلمات اضافی هر متن حذف، ریشهٔ کلمات استخراج و کلمات اصلی متن وارد پایگاه داده متن شود. برای انجام این‌کار، فایلی شامل کلمات اضافی متداول درنظر گرفته شده است. متن مورد نظر کلمه‌به‌کلمه خوانده شده و هر کلمه با کلمات موجود در فایل کلمات اضافی مقایسه می‌شود؛ درصورتی‌که کلمه مورد نظر در فایل کلمات اضافی وجود داشته باشد، از متن حذف و در غیراین‌صورت به رکورد متن مورد نظر در بانک اطلاعاتی پایگاه داده اضافه می‌شود. به این‌ترتیب با یک پویش ترتیبی، کلمات اضافهٔ رایج در متن حذف می‌شوند.

۲-۲-۴- مرحله دوم پیش پردازش

در مرحله دوم پیش‌پردازش با به‌کارگیری فرمول ساده‌ای برخی پیشوندها و پسوندهای متداول از کلمات باقی مانده جدا شده و سعی می‌شود ریشه کلمات باقی بماند. برای مثال کلمه‌ای مانند "درختان" که "ان"، پسوند جمع برای آن به حساب می‌آید، از کلمه درخت حذف می‌شود. قابل توجه است که به سبب وجود استثناهای فراوان در کلمات فارسی، انجام مرحله حذف پیشوندها و پسوندها همواره به‌سادگی امکان‌پذیر نیست. برای مثال برای کلمه‌ای مثل "باران" چون "ان" بخش اصلی کلمه است، امکان حذف نیست. درنتیجه در این مرحله به حذف پیشوندها و پسوندهای کلمات که حرف اصلی آن بیش از سه حرف باشد، اقدام کرده‌ایم. مشاهدات و بررسی‌ها نشان داده است که در بسیاری از موارد، درنظرگرفتن چنین معیاری منجر به حذف صحیح پیشوند و پسوند می‌شود. با این حال از روش‌های پیشرفته‌تری به‌منظور استخراج ریشه از کلمات اصلی می‌تواند استفاده شود. در جداول (۴ و ۵) برای هر دسته موضوعی نشان داده شده که عملیات پیش‌پردازش متن به‌طور متوسط چه حجمی از کلمات را بیرون ریخته و چه درصدی را به‌عنوان کلمات اصلی به کاربر ارائه می‌کند.

(جدول-۴): فایل‌های پالایش‌شده

(Table-4): Refined files

ردیف	دسته بندی موضوعی	متوسط تعداد کلمات مقالات	متوسط تعداد کلمات پس از حذف کلمات اضافی وریشه یابی
۱	ورزشی	۲۰۴	۱۴۹
۲	اقتصادی	۱۹۹	۱۳۵
۳	شهری	۱۲۳	۷۶
۴	حوادث	۱۶۰	۱۱۵
۵	خارجی	۱۷۷	۱۲۴

(جدول-۵): نتایج

(Table-5): Results

ردیف	دسته‌بندی موضوعی	درصد کلمات اضافی که به اشتباه جزو کلمه اصلی شناسایی شدند	درصد کلمات اصلی که به اشتباه جزو کلمه اضافی شناسایی شدند
۱	ورزشی	۲۰٪	۹٪
۲	اقتصادی	۱۸٪	۱۱٪
۳	شهری	۲۲٪	۱۲٪
۴	حوادث	۱۶٪	۱۴٪
۵	خارجی	۱۵٪	۱۰٪

[1] Site address: http://www.hamshahrionline.ir

۳-۲-۴- مرحله یافتن تعداد کلمات اصلی موجود درمتن

پس از انجام مراحل پیش‌پردازش متن که طی آن کلمات اصلی متن استخراج شدند، نوبت به شمارش تعداد تکرار کلمات موجود در متن می‌رسد. در روش‌هایی که تابه‌حال در زبان فارسی به این منظور استفاده شده، هم‌خانواده‌بودن و وجود کلمات اعم و اخص در یک متن مورد توجه قرار نمی‌گرفته است. همان‌گونه که گفته شد از آن‌جایی‌که نویسندگان متن، از کلمات هم‌خانواده برای شکیل‌بودن متن استفاده کنند، عدم درنظرگرفتن هم خانواده‌ها و کلمات اعم و اخص باعث کاهش دقت فرایند طبقه‌بندی متن می‌شود. به‌عبارت دیگر چون یک کلمه به شکل‌های مختلفی در متن به‌کار رفته است، تعداد تکرار کلمه مورد نظر درمتن توزیع می‌شود. درنتیجه شمارش ساده معیار مناسبی برای تعیین طبقه متن یا موضوع متن نیست. با استفاده از یک اصطلاح‌نامه به‌سادگی کلمات هم‌خانواده و اعم و اخص را می‌توان شناسایی کرده و برای همه این کلمات یک نماینده در متن در نظر گرفته شود و با دیدن کلمات هم‌خانواده و اعم و اخص در متن به میزان ضریب وزنی تعریف‌شده به فرکانس تکرار کلمه نماینده اضافه شود. برای پیاده‌سازی این مرحله از یک اصطلاح‌نامه کامل متون فارسی (اصفا) استفاده شده است که در فصل قبل به‌طور کامل مورد بررسی قرار گرفته است.

۳-۴- بررسی ساختار اصطلاح‌نامه

این ساختار در یک فایل اکسس قابل نمایش است که در شکل (۶) قسمتی از آن نشان داده می‌شود. ساختار اصطلاح‌نامه از ترکیبی از ساختار اصطلاح‌نامه اسلامی و اصفا گرفته شده و داده‌های آن از پایگاه کتابخانه ملی استخراج شده و پس از تبدیل به فرمت لازم برگردانده شده است. لازم به ذکر است که تهیه این اصطلاح‌نامه بسیار سخت بوده و قبل از به‌دست‌آوردن این اصطلاح‌نامه از اصطلاح‌نامه‌های زیادی استفاده شد که به‌علت کامل‌نبودن نتیجهٔ خوبی به‌دست نیامد و تنها این اصطلاح‌نامه بود که نتایج خوبی از آن منتج شد.

۴-۴- ایجاد بردارها

پس از مشخص‌شدن تعداد تکرارهای کلمات موجود در متن با درنظرگرفتن هم‌خانواده‌ها، اعم و اخص‌های کلمات، نوبت به ایجاد بردار می‌رسد. بدین‌منظور برای هر متن یک بردار در نظر گرفته می‌شود که حاوی مقادیر صفات است. صفات در این‌جا کلیه کلمات اصلی موجود در بانک داده‌ای متنها هستند. بدین‌ترتیب هر عنصر از یک بردار نماینده، یک کلمه اصلی بوده و برای یک متن خاص درصورت دارابودن کلمات مورد نظر، عنصر مربوطه در بردار، مقدار تکرار کلمه را گرفته و درصورت عدم حضور کلمه در متن مورد نظر، مقدار عنصر مورد نظر صفر خواهد بود. نمونه‌ای از یک بردار تهیه‌شده برای یک متن ذکرشده در شکل (۷) قابل مشاهده است. با ساخت بردارهای مربوط به متن‌های مختلف با استفاده از روش KNN متون موجود در بانک طبقه‌بندی می‌شود.

terms | term_relations

keyword_id	keyword	type	keyword_en
28	رقص های آیینی	K	Ritual Dances
29	رقص جنگ	K	War Dance
30	آداب معاشرت	K	Etiquette
31	مراسم خاک سپاری	K	Burial Ceremor
32	آداب و رسوم مذهبی	K	Religious Custc
33	حج	K	Pilgrimage to M
34	نماز[۱]	K	Prayer
35	نماز جماعت	K	Public Player
36	روزه	K	Fasting
37	زیارت[۱]	K	Pilgrimage
38	باران خواهی	K	Praying for Rai
39	درون پسن	K	
40	زوهر	K	
41	پسنا خوانی	K	
42	نذر	K	Offering
43	طهارت	K	Purification

terms | term_relations

keyword_id	parent_keyv	relation_type
18	20	B
21	22	B
21	23	B
24	25	B
24	27	B
25	26	B
28	29	B
32	33	B
32	34	B
32	36	B
32	38	B
32	39	B
32	40	B

(شکل-۶): ساختار اصطلاح‌نامه

(Figure-6): The structure of the thesaurus

V1	V2	V3	V4	V5	V6	....	V1256	V1256	....
0	2	0	1	8	3	....	2	4	....

(شکل-۷): نمونه‌ای از یک بردار متن

(Figure-7): An example of a text vector

۵-۴- وزن‌دهی کلمات اصلی

جهت ادامه روند کار و اجرای روش پیشنهادی، ضمن بررسی

راه‌کارهای ممکن، از یک الگوریتم که در مقاله‌ای با عنوان Text Relatedness Based on a Word Thesaurus توسط آقای George Tsatsaronis [36] ارائه شده بود و توضیحات آن در بخش‌های قبلی ارائه شد استفاده کردیم. در این مقاله پژوهش‌گر به‌دنبال به‌دست‌آوردن وزن مناسب با توجه به ارتباط معنایی بین کلمات است که بر اساس این اوزان، مناسب‌ترین شاخص‌های کلیدی را در متون بتواند استخراج کند. الگوریتم ارائه‌شده برای دسته‌بندی متون لاتین استفاده شده بود که در این مقاله از آن با پاره‌ای تغییرات برای زبان فارسی استفاده شده است. پس از یافتن شاخص‌های کلیدی متون به‌دست آمده می‌بایست طبقه‌بندی شوند که به این منظور از الگوریتم نزدیکترین فاصله KNN استفاده شده است. از آن‌جایی‌که در مرحله نخست هدف یافتن میزان تکرار کلمات هم‌خانواده در متن است، با شروع از متن کلمات اصلی، برای هر کلمه اصطلاح‌نامه مورد نظر مورد بررسی قرار می‌گیرد؛ درصورتی‌که کلمه خوانده‌شده هم‌خانواده کلمهٔ دیگری باشد که قبلاً در متن خوانده شده بود، به‌جای اضافه‌کردن تکرار کلمه جدید خوانده شده، درصد وزنی معادل ضریب Beta (که بین صفر تا یک است) به شمارنده کلمه نخست نماینده اضافه می‌شود. هم‌چنین از آن‌جایی که برای کلمه نماینده اعم و اخص، وزنی مساوی کلمات تکرار کلمه نمی‌توان در نظر گرفت با دیدن اعم و اخص یک کلمه، به وزن کلمه نماینده درصد وزنی مشخصی معادل Alpha اضافه می‌شود (جلوتر در بخش نتایج، بر اساس آزمایش‌های بسیار زیاد و مقایسه‌های زیادی که برای به‌دست‌آوردن مقدار Alpha و Beta به‌صورت بهینه و مناسب انجام‌شده به این نتیجه خواهیم رسید که در حالتی مقدار Alpha=0.2 و Beta=0.4 باشد عملکرد الگوریتم در یافتن شاخص‌های کلیدی بهترین حالت است)؛ اما درحقیقت با یک پویش یعنی خواندن از نخست متن تا آخر متن ویژگی‌ها (کلمات کلیدی) را نمی‌توانیم مدیریت و بردار ویژگی‌ها را استخراج کنیم (الگوریتم حریصانه).

برای این که ویژگی‌ها را مدیریت کنیم دو روش وجود دارد؛ روش نخست این است که به‌ازای هر کلمه باید برگردیم و تمام کلمات کلیدی را بررسی کنیم و برای متونی مناسب است که تعداد کلمات کلیدی‌شان کم باشد. به‌عنوان مثال فرض می‌کنیم، نخستین کلمه‌ای که در متن دیدیم، گرگ است و دومین کلمه جانور. الان وزن گرگ یک است و وزن جانور به‌دلیل این‌که اخصش یعنی گرگ را دیدیم $\beta + \alpha$ می‌شود. حال گرگ هم به‌خاطر این‌که اعمش یعنی جانور را دیدیم $\beta + \alpha$ می‌شود. اگر کاینات هم در متن بیاید و چون جانور نوعی از کاینات است و گرگ نوعی جانور، پس وزن کاینات $1 + \alpha + \alpha^2$ می‌شود و به وزن کلمه‌ی جانور یک α دیگر اضافه می‌شود و داریم $(\beta + \alpha) + \alpha$ و به وزن گرگ هم به‌دلیل این که اخص آن کاینات است یک α^2 اضافه می‌شود و داریم $(\beta + \alpha) + \alpha^2$. پس نتیجه می‌گیریم یک الگوریتم حریصانه نمی‌توانیم ارائه دهیم چون وقتی یک کلمه می‌بینیم، ممکن است بخواهیم وزن کلمات بعدی رو تغییر دهیم.

روش دوم این است که ابتدا بردار ویژگی‌ها را به‌دست آوریم و بعد آن را روی کلمات کلیدی اعمال کنیم. یعنی یک‌بار تا آخر متن می‌رویم؛ سپس کلمات کلیدی را روی بقیه ویژگی‌ها پخش می‌کنیم؛ که در این روش سربار زمانی خیلی کمتر است.

در روش نخست به‌ازای هر کلمه باید برمی‌گشتیم همهٔ کلمات کلیدی را چک می‌کردیم؛ بنابراین سربار زمانی زیاد بود. به‌عنوان مثال اگر کلمه کلیدی جانور سیزده بار دیده شده باشد و گرگ یازده بار، درنتیجه برای وزن‌دهی $11 * \alpha$ بعلاوه‌ی جانور و $13 * \alpha$ به‌علاوه‌ی گرگ می‌شود؛ پس اگر کلمات کلیدی زیاد باشد، بهتر است که نخست همهٔ آن‌ها را استخراج کنیم و بعد یک کلمه را به سایر کلمات انتشار بدهیم. مسأله دیگری که مطرح می‌شود، این است که به‌عنوان مثال ممکن است، چهار کلمهٔ شیر، پلنگ، گرگ و خرس را در متن ببینیم که یکی چهار بار، یکی سه بار، یکی دوبار و دیگری یک بار تکرار شده است و کلمهٔ دیگری به‌عنوان اعم یا اخص در متن نباشد؛ پس لازم است کلمهٔ جانور را به متن اضافه کنیم. در غیر این صورت ممکن است با متن دیگری که فقط جانور در آن بوده تطابق پیدا نکند. درحقیقت اگر وزن از یک حدی گذشت باید خودبه‌خود کلمهٔ عام‌تر (جانور) تولید شود. بنابراین باید یک سطح آستانه بگذاریم که ویژگی‌ها یا کلمات کلیدی‌هایی که از آن سطح آستانه بیشترند، ظاهر شوند؛ یعنی این چهار کلمه که کلمات خاصی از کلمهٔ جانور هستند باعث می‌شوند کلمهٔ جانور که اصلاً در متن نبود ظاهر شود. اگر این سطح آستانه را نگذاریم، همهٔ کلمات تولید می‌شوند و به‌محض این‌که کلمه‌های اخص را دیدیم، هر کدام با وزنی ظاهر می‌شوند یکی ۰,۲۵ یکی ۰,۰۶۲۵ و ... ؛ پس بعد از گذاشتن سطح آستانه کلمات کلیدی را که زیر آن هستند، حذف می‌کنیم. به این‌ترتیب با یک پویش ترتیبی کلمات اصلی مشخص می‌شود که آیا از قبل هم‌خانواده آن کلمه در متن دیده شده

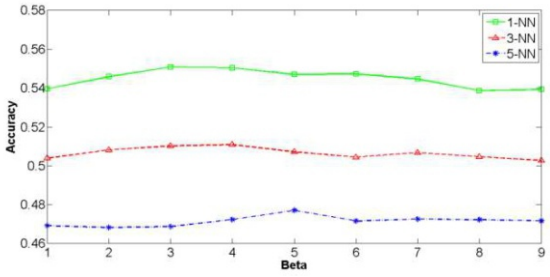
بود، که در این‌صورت درصد وزنی 0.4 به شمارنده کلمه نماینده اضافه می‌شود و اگر اعم یا اخص کلمه مورد نظر بود، مقدار 0.2 به شمارندهٔ کلمهٔ نماینده اضافه می‌شود.

۶-۴- اجرای الگوریتم و تجزیه تحلیل نتایج

هدف از اجرای الگوریتم این است که بهترین ضریبی که برای پارامترهای Alpha (α) و Beta (β) می‌تواند در نظر گرفته شود، مشخص شوند. برای این منظور، ضریب‌های مختلفی (بین صفر تا یک) با دقت و فاصله 0.1 برای هر پارامتر در نظر گرفته و در هر بار الگوریتم پیاده‌سازی‌شده به‌ازای این مقادیر اجرا می‌شود و درنهایت کلیه پاسخ‌های به‌دست‌آمده مورد مقایسه و ارزیابی قرار می‌گیرد تا درنهایت به یک مقدار مناسب برای هر یک از پارامترها دست یافت. البته به‌صورت منطقی می‌توان حدس زد که مقداری که برای ضریب هم‌خانوادگی محاسبه می‌گردد، بیشتر از ضریبی باشد که برای کلمات اعم و اخص حاصل می‌شود. زیرا کلمات هم‌خانواده از یک ریشه به‌وجود می‌آیند و به‌طورعمومی این‌گونه کلمات دارای بار موضوعی مشابه‌تری نسبت به کلمات اعم و اخص دارند. با این حال دست‌یابی به میزان دقیق این پارامترها از طریق اجرای الگوریتم امکان‌پذیراست. معیار ارزیابی در این الگوریتم میزان دقت حاصل از اجرای برنامه است.

۱-۶-۴- بهترین دقت برای Beta

با توجه به شکل (۸) روش کار به این‌صورت است که پارامتر β از ۰,۱ تا ۰,۹ تغییر داده شده است و برای هر کدام از خانه‌های β (به‌عنوان مثال خانهٔ نخست)، α نیز از ۰,۱ تا ۰,۹ تغییر داده شده است؛ بنابراین هر بار یک دقت به‌دست می‌آید. به‌عنوان مثال میانگین همه دقت‌ها در رده‌بند $5-NN$ برابر 0.47 شد. همین‌کار برای رده‌بندهای $1-NN$ و $3-NN$ نیز انجام می‌شود. پس در نتیجه ۹ تا α گوناگون با $\beta = 0.2$، ۹ تا α گوناگون با $\beta = 0.3$، ۹ تا α گوناگون با $\beta = 0.4$ و ... داریم. درنتیجه هر کدام از رده‌بندها ۸۱ بار میانگین گرفته شده است.



(شکل-۸): دقت روابط هم‌معنی

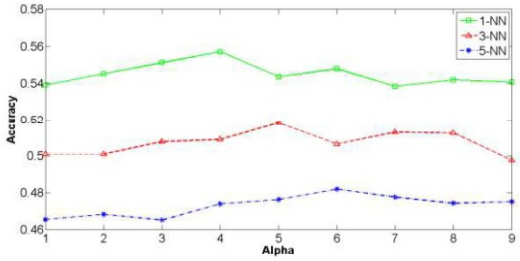
(Figure-8): meaning relationship accuracy

میانگین نتایج برای α از 0.1 تا 0.9 با گام 0.1 ارائه شده است.

۱. از این شکل می‌فهمیم که $1-NN$ بهتر از دیگر طبقه‌بندها است.

۲. از این شکل می‌فهمیم که بهترین نقطه برای β، مقدار 0.3 است.

در شکل (۹) که برای $\beta = 0.3$ تنظیم شده است، β برابر 0.3 در نظر گرفته شده است و α را تغییر داده‌ایم. یعنی $\alpha = 0.2, \beta = 0.3$، $\alpha = 0.3, \beta = 0.3$، $\alpha = 0.4, \beta =$ 0.3 و ... . پس در این شکل β تغییر نمی‌کند، در نتیجه فقط نُه بار میانگین گرفتیم.



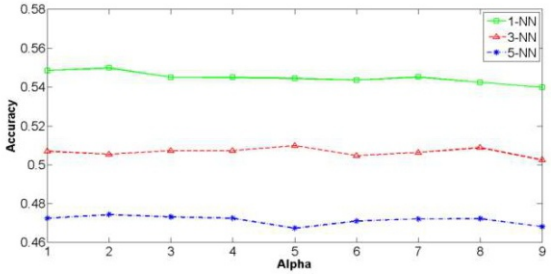
(شکل-۹): بهترین دقت (۱)

(Figure-9): Best accuracy (1)

همان‌طورکه از شکل پیداست با تنظیم $\beta = 0.3$ و $\alpha =$ 0.4 بهترین دقت به‌دست می‌آید.

۲-۶-۴- بهترین دقت برای Alpha

این‌بار بهترین دقت را برای میزان مشارکت روابط اعم و اخص می‌خواهیم به‌دست آوریم. روش کار همانند به‌دست‌آوردن دقت برای β است با این تفاوت که این بار α را از 0.1 تا 0.9 تغییر می‌دهیم و برای هر کدام از خانه‌های α (به‌عنوان مثال خانهٔ نخست)، β را از 0.1 تا 0.9 تغییر می‌دهیم؛ بنابراین هر بار یک دقت به‌دست می‌آید.



(شکل-۱۰): دقت روابط اعم و اخص

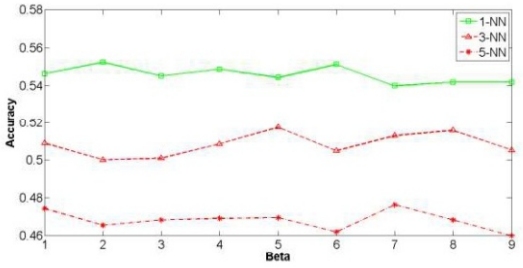
(Figure-10): accuracy of the general and specific relationship

با توجه به شکل میانگین نتایج برای β از 0.1 تا 0.9 با گام 0.1 ارائه شده است.

۱. از شکل (۱۰) می‌فهمیم که 1-NN بهتر از دیگر طبقه‌بندها است.

۲. از این شکل می‌فهمیم که بهترین نقطه برای α، 0.2 است.

در شکل (۱۱) که برای $\alpha = 0.2$ تنظیم شده است و β را تغییر داده‌ایم. یعنی $\alpha = 0.2, \beta = 0.2$، $\alpha = 0.2, \beta = 0.3$، $\alpha = 0.2, \beta = 0.4$ و ... . پس در این شکل α تغییر نمی‌کند، در نتیجه فقط نُه بار میانگین گرفتیم.

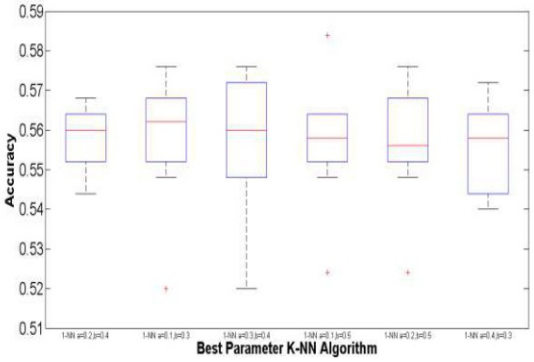


(شکل-۱۱): بهترین دقت (۲)
(Figure-11): Best accuracy (2)

همان‌طور که از شکل (۱۱) پیداست با تنظیم $\alpha = 0.6$,$\beta = 0.2$ بهترین دقت به‌دست می‌آید.

۳-۶-۴- بهترین دقت‌ها برای مدل K-NN

در شکل (۱۲)، نتایج شش تا از بهترین کارهایی که ثبت شده، ارائه شده است.

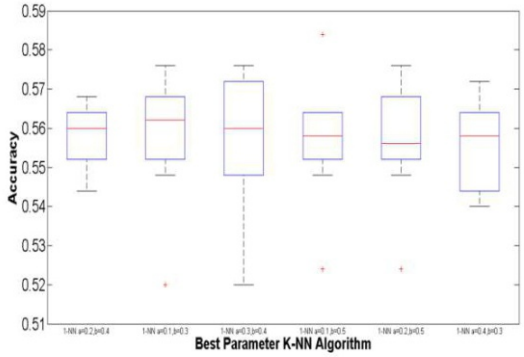


(شکل-۱۲): بهترین دقت‌ها برای مدل K-NN
(Figure-12): Best accuracy for the K-NN model

این نتایج، α و βهای گوناگونی هستند که بهترین دقت‌ها را به‌دست آوردند. به این‌صورت که ۸۱ حالت را آزمایش کرده‌ایم (نُه حالت، α از 0.1 تا 0.9 و نُه حالت β از 0.1تا 0.9) و شش‌تای برتر را در این‌جا آورده‌ایم.

همان‌گونه که عنوان شد، مهم‌ترین پارامتر در کسب نتیجهٔ بهتر، دارابودن دقت بالاتر در ارزیابی و تجزیه و تحلیل نتایج هر مرحله است؛ لذا به‌منظور انجام ارزیابی، الگوریتم اعتبارسنجی ده لایه به‌کار گرفته شده است. در این الگوریتم نود درصد از متون داده‌ای، به‌عنوان آموزش و تعلیم در اختیار الگوریتم قرار داده می‌شود و ده درصد بقیه جهت آزمایش الگوریتم در نظر گرفته شده است؛ سپس با درنظرگرفتن داده‌های آموزشی مقدار مؤلفه‌ها و متغیرهای اصلی حدس زده می‌شود. جهت آزمایش و تأیید نتایج حاصله، از داده‌های آزمون؛ استفاده می‌شود. در این ارزیابی البته برای دست‌یابی به نتیجه بهتر و افزایش دقت و حساسیت در کار، به‌ازای هر پارامتر، ده بار الگوریتم اجرا می‌شود و متوسط نتایج در نظر گرفته می‌شود. پارامترهایی که نتایجی بهتر و نزدیک‌تر به واقعیت ایجاد کنند، مورد نظر خواهند بود. همچنین برای بیان نحوه توزیع داده‌ها در هر مرحله از نمودار جعبه‌ای توزیع استفاده شده است (هر چه جعبه در نمودار کوچک‌تر و دادهٔ مجزای کم‌تری داشته باشد، نتیجه بهتر خواهد بود).

با توجه به توضیحات ارایه‌شده در نمودار جعبه‌ای شکل (۱۳)، به‌ازای شش مورد از بهترین α و βهای به‌دست‌آمده الگوریتم را ده بار اجرا کرده ونتایج به‌دست آمده نشان داده شده است. همان‌گونه که مشاهده می‌شود در $\alpha = 0.2$ و $\beta = 0.4$ جعبه نمایان‌گر اجراهای منسجم‌تر بوده و بنابراین این نقطه بهترین نقطه از نظر این پژوهش خواهد بود.



(شکل-۱۳): نمودار جعبه‌ای برای الگوریتم ارائه شده
(Figure-13): Box plot for the proposed algorithm

جدول (۶) دقت طبقه‌بندی را برای حالت استفاده از اصطلاح‌نامه و بدون اصطلاح‌نامه نشان می‌دهد.

(جدول-۶): مقایسه خروجی‌ها از نظر دقت
(Table-6): Compare outputs in terms of accuracy

روش‌ها	عدم استفاده از اصطلاح‌نامه (β=0 و α=0)	استفاده از اصطلاح‌نامه، تنها در سطح مترادف (β=1 و α=0)	استفاده از اصطلاح‌نامه با روش پیشنهادی (α=0.25 و β=0.3)
1 – NN	۷۰/۹۳	۷۳/۲۱	**۷۳/۶۸**

$3-NN$	۶۹/۹۱	۷۳/۰۷	**۷۳/۱۰**
$5-NN$	۷۰/۵۷	۷۳/۱۴	**۷۳/۶۱**
MLP	۶۸/۰۲	۷۰/۸۴	**۷۱/۲۸**

همان‌گونه که از جدول (۶) پیداست استفاده از اصطلاح‌نامه با توجه به روش پیشنهادی به طبقه‌بندی دقیق‌تر متون فارسی می‌تواند کمک کند. هم‌چنین، چنان‌چه از دقت‌ها معلوم است، مدل KNN از مدل MLP بهتر است. هم‌چنین در بین KNN-های گوناگون، با افزایش K دقت بدتر می‌شود؛ یعنی بهترین مدل $1-NN$ است.

۴-۶-۴- مقایسهٔ دقت روش پیشنهادی با سایر روش‌ها

در این بخش روش پیشنهادی با چندین روش دیگر از نظر دقت مورد مقایسه قرار می‌گیرد. نتایج حاصل‌شده از جدول (۷) حاکی از برتری روش پیشنهادی نسبت به سایر روش‌های مشابه است.

(جدول-۷): مقایسه روش پیشنهادی و سایر روش‌ها
(Table-7): Comparison of the proposed method and other methods

روش‌ها	دقت
روش پیشنهادی در بهترین حالت	۷۳/۶۸
روش نزدیکترین همسایه یغمایی و تعبدی [3]	۶۷/۳۸
روش آراسته و همکارانش [5]	۶۴/۶۳
روش علاقه‌بند و همکاران [4]	۶۷/۶۴

۵- نتیجه‌گیری و کارهای آینده

امروزه با افزایش روزافزون حجم اطلاعات، وجود سامانه‌ای برای طبقه‌بندی خودکار متون ضروری به‌نظر می‌رسد. در این مقاله سامانهٔ جدیدی برای طبقه‌بندی خودکار متون فارسی ارائه شد. این سامانه شامل چهار مرحلهٔ اصلی است: مرحلهٔ پیش‌پردازش، مرحلهٔ استفاده از اصطلاح‌نامه، مرحلهٔ وزن‌دهی و سپس مرحله طبقه‌بندی. در مرحله وزن‌دهی با پردازش داده‌های آموزشی بهترین ویژگی‌های نماینده هر طبقه استخراج شده و برای آموزش طبقه‌بندی‌کننده مبتنی بر KNN استفاده می‌شوند و سپس در مرحلهٔ طبقه‌بندی، KNNقادر خواهد بود داده‌های آزمایش را به یکی از طبقه‌های آموزش داده شده نسبت دهد.

در این مقاله نکات زیر مورد توجه قرار گرفته است: با استفاده از اصطلاح‌نامه کلمات وابسته در متون فارسی شامل کلمات هم‌خانواده، کلمات اعم و اخص شناسایی شده‌اند. با استفاده از ارتباط بین کلمات موجود در متن با استفاده از اصطلاح‌نامه، وزن‌دهی دقیق انجام شده است و به این‌ترتیب شاخص کلیدی با دقت بیشتری استخراج می‌شوند. با تبدیل هر متن به یک فضای برداری که ابعاد آن شاخص‌های کلیدی همه متون هستند، عمل طبقه‌بندی متون بر اساس شاخص‌های استخراج‌شده انجام می‌شود. نتایج حاصل از به‌کارگیری اصطلاح‌نامه با توجه به روش پیشنهادی در استخراج کلمات کلیدی متون فارسی بر طبقه‌بندی آن‌ها با حالت بدون اصطلاح‌نامه و استفاده از اصطلاح‌نامه فقط در سطح مترادف مقایسه شد. نتایج بررسی‌شده نشان داد که استفاده از اصطلاح‌نامه با توجه به روش پیشنهادی به دسته‌بندی دقیق‌تر متون فارسی می‌تواند کمک کند. با توجه به پیاده‌سازی و اجرای الگوریتم ارائه‌شده و نتایج به‌دست‌آمده از تجزیه و تحلیل خروجی‌های الگوریتم این‌گونه به‌نظر می‌رسد که، چنان‌چه برای کلمات و واژه‌های دارای روابط اعم و اخص ضریب وزنی 0.2 و هم‌چنین برای کلمات دارای روابط هم خانواده ضریب وزنی 0.4 در نظر گرفته شود، شواهد به‌دست آمده از آزمایش داده‌ها و متون آزمایشی بیان‌گر انتخاب مناسب این پارامترها در تشخیص کلمات کلیدی و درنهایت دسته‌بندی متون می‌باشد. البته به‌صورت منطقی هم نتایج به‌دست آمده از الگوریتم که بیان‌کننده مقادیر ضریب وزنی است، قابل قبول است؛ زیرا کلمات هم‌خانواده که دارای ریشه مشترک هستند، به‌طورعمومی در یک موضوع و دسته خاص مورد استفاده قرار می‌گیرند و به‌همین دلیل ضریب بالاتری نیز خواهند داشت.

۵-۱- پیشنهادهای آینده

موارد زیر را به‌عنوان پیشنهاد جهت پروژه‌های آینده می‌توان مطرح کرد:

۱- بهبود مرحله پیش‌پردازش؛
۲- استفاده از روش‌های دیگر طبقه‌بندی؛
۳- ارائه روش‌های مناسب‌تری جهت ابهام‌زدایی از کلمات؛
۴- تعیین بهترین ضریب برای روابط متضاد و همبسته؛
۵- به‌کارگیری روابط دیگر اصطلاح‌نامه.

6-References ۶- مراجع

[۱] راد، ف.، پروین، ح.، دهباشی، آ.، مینایی، ب.، ارائه روشی جدید برای شاخص‌گذاری خودکار و استخراج کلمات کلیدی برای بازیابی اطلاعات و خوشه‌بندی متون، نشریه پردازش علائم و داده‌ها، پیاپی ۲۷، شماره ۱، صفحه ۸۷-۱۰۰، ۱۳۹۵.

[1] F. Rad, H. Parvin, A. Dehbashi, B. Minaei, "A New Method for Automatic Indexing and Extracting Keywords for Information Retrieval and Clustering of Texts", Journal of Signal Processing and Data, Volume 27, No. 1, page 87-100, 2017.

[۲] دهباشی هاشم، آتوسا، بهبود خوشه بندی متون فارسی بر اساس کلمات کلیدی با استفاده از اطلاعات زبان شناختی و اصطلاح‌نامه. پایان‌نامه کارشناسی ارشد، ۱۳۸۹.

[2] Dehbashi Hashem, Atoosa, "Improved clustering of Persian texts based on keywords using linguistic information and thesaurus". Master thesis, 2010.

[۳] یغمایی، ف.، تعبدی س، بهبود دسته‌بندی متون فارسی در روش همسایگی وزن‌دار، نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی، ۱۳۹۱.

[3] F. Yaghmaei, S. Tabodi, "Improving the Classification of Persian Texts in Weighted Neighboring Method", The First International Conference on Line Processing and Persian Language, 2012.

[۴] علاقه بند، م.ر.، سعیدی محمدی، م.ر.، دزفولیان، م.ح.، خوشه‌بندی متون مبتنی بر مرکز دسته با استفاده از روش SVD و بهره‌گیری از نقاط همسایگی، نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی، ۱۳۹۱.

[4] M.R., Alagheband, M.R Saeedi Mohammadi, M.H Dezfulian, "clustering of center-based texts using the SVD method and utilizing neighborhoods", the first international conference on processing Persian language and language, 2012.

[5] A.R, Arasteh, M.H, Elahimanesh, A. Sharif, B. Minaei-Bidgoli, "Semantically Clustering of Persian Words", Proceeding of 1st International Conference on Persian Language Processing (ICPLP), Semnan, Iran, Sep. 5-6, 2012.

[6] Berry, W. Michael, and Castellanos, Malu, eds, Survey of text mining. New York: Springer, 2004.

[7] Borko, Harold, and Bernick, Myrna, "Automatic document classification", Journal of the ACM (JACM) 10, no. 2: 151-162, 1963.

[8] Cavnar, B. William, and Trenkle, M. John, "N-gram-based text categorization", Ann Arbor MI 48113, no. 2: 161-175, 1994.

[9] F. Colace, M. D. Santo, L. Greco, P. Napoletano, "Text classification using a few labeled examples", Journal of Computers in Human Behavior, Vol. 30, January 2014, pp. 689-697, 2014.

[10] Cleverdon, Cyril, "Optimizing convenient online access to bibliographic databases', Information services and Use 4, no. 1: 37-47, 1984.

[11] D. Choi, B. Ko, H. Kim, P. Kim, Text analysis for detecting terrorism-related articles on the web, Journal of Network and Computer Applications, Vol. 38, pp. 16-21, 2014.

[12] A. Díaz, M. Buenaga, L. A. Ureña, and M. García, "Integrating Linguistic Resources in an Uniform Way for Text Classification Tasks", In First International Conference on Language Resources & Evaluation, Granada (Spain), 1998.

[13] M. Deegan, "Keyword Extraction with Thesauri and Content Analysis", URL: http://www.rlg.org/en/page.php?Page_ID=17068, 2004.

[14] Escudero, Gerard, Màrquez, Lluís, and Rigau, German, "Boosting applied to word sense disambiguation", Springer Berlin Heidelberg, 2000.

[14] K. Frantzi, S. Ananiadou and H. Mima, Automatic Recognition of Multi-word Terms: the C-value/NC-value Method, Digital Libraries, 3(2), pp. 115-130, 2000.

[15] S. Forsyth, Richard, "New directions in text categorization", In Causal models and intelligent data management, pp. 151-185. Springer Berlin Heidelberg.

[16] N. Freitas, and A. Kaestner, "Automatic text summarization using a machine learning approach", 16th Brazilian Symposium on Artificial Intelligence (SBIA), Brazil. Vol. 398, 2005.

[17] Granitzer, Michael, Hierarchical text classification using methods from machine learning. Master's Thesis, Graz University of Technology, 2003.

[18] D. Hyun, "Automatic Keyword Extraction Using Category Correlation of Data", Heidelberg, pp. 224-230, 2006.

[19] Harter, Stephen P. "A probabilistic approach to automatic keyword indexing", Part II. An algorithm for probabilistic indexing. Journal of the American Society for Information Science 26, no. 5: 280-289, 1975.

and Applications, pp. 391-398. Springer Berlin Heidelberg, 2011.

[32] Sable, L. Carl, and Hatzivassiloglou, Vasileios. "Text-based approaches for non-topical image categorization", International Journal on Digital Libraries 3, no. 3: 261-275, 2000.

[33] Salton, Gerard, and Yang, Chung-Shu, "On the specification of term values in automatic indexing", Journal of documentation 29, no. 4: 351-372, 1973.

[34] Schapire, E. Robert, and Singer, Yoram, "BoosTexter: A boosting-based system for text categorization", Machine learning 39, no. 2-3: 135-168, 2000.

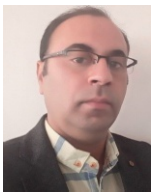
[35] G. Tangil, J. E. Tapiador, P. Peris-Lopez, J. Blasco, Dendroid: A text mining approach to analyzing and classifying code structures in Android malware families, Journal of Expert Systems with Applications, Vol. 41, No. 4, March 2014, pp. 1104-1117, 2014.

[36] G. Tsatsaronis, I. Varlamis, M. Vazirgiannis, "Text Relatedness Based on a Word Thesaurus", Journal of Artificial Intelligence Research, Vol. 37 pp.1-39, 2010.

[37] A. Zamanifar, B. Minaei-Bidgoli, and Sharifi, Mohsen. "A new hybrid farsi text summarization technique based on term co-occurrence and conceptual property of the text. Software Engineering, Artificial Intelligence", Networking, and Parallel/Distributed Comput-ing, SNPD'08. Ninth ACIS International Conference on. IEEE, 2008.

[38] W. Witten, I.H. Medley, Thesaurus based automatic keyphrase indexing, ACM/IEEE-CS JCDL '06 (Joint Conference on Digital Libraries), 2006.

[39] Y. Zhang, N. Z. Heywood and E. Milios, "World Wide Web Site Summarization Web Intelligence and Agent Systems", Technical Report, 2006.



مجید محمدپور دانش‌آموخته کارشناسی ارشد رشته نرم‌افزار از دانشگاه علوم تحقیقات تهران (کهگیلویه و بویراحمد) در سال ۱۳۹۳ است. وی هم‌اکنون در چندین واحد دانشگاهی در رشته کامپیوتر مشغول به تدریس است. ایشان هم‌اکنون عضو باشگاه پژوهش‌گران جوان و نخبگان دانشگاه آزاد اسلامی واحد یاسوج و در چندین واحد دانشگاهی در رشته کامپیوتر مشغول به تدریس است.

[20] Hassel, Martin, and Mazdak, Nima, FarsiSum: a Persian text summarizer. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. Association for Computational Linguistics, 2004.

[21] Huang, Yan. "Support vector machines for text categorization based on latent semantic indexing", Electrical and Computer Engineering Department, The Johns Hopkins University, Tech. Rep, 2003.

[22] Kessler, Brett, Numberg, Geoffrey, and Schütze, Hinrich, Automatic detection of text genre. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, pp. 32-38. Association for Computational Linguistics, 1997.

[23] Knight, Kevin, Mining online text. Communications of the ACM 42, no. 11: 58-61, 1999.

[24] Larkey, S, Leah, "Automatic essay grading using text categorization techniques", In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 90-95. ACM, 1998.

[25] Liu, Luying, Kang, Jianchu, Yu, Jing and Wang. Zhongliang, "A comparative study on unsupervised feature selection methods for text clustering. In Natural Language Processing and Knowledge Engineering", 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on, pp. 597-601. IEEE, 2005.

[26] H. P, Luhn, "11 Keyword-in-Context Index for Technical Literature (KWIC Index)", Readings in automatic language processing 1: 159, 1996.

[27] Manning, D. Christopher, "Foundations of statistical natural language processing", Edited by Hinrich Schütze. MIT press, 1999.

[28] Maron, Melvin Earl., "Automatic indexing: an experimental inquiry", Journal of the ACM (JACM) 8, no. 3: 404-417, 1961.

[29] Myers, Kary, Kearns, Michael, Singh, Satinder, and Walker, A. Marilyn, "A boosting approach to topic spotting on subdialogues", Family Life 27, no. 3: 1, 2000.

[30] Moschitti, Alessandro, "Answer filtering via text categorization in question answering systems", In Tools with Artificial Intelligence, Proceedings. 15th IEEE International Conference on, pp. 241-248. IEEE, 2003.

[31] H. Parvin, B. Minaei-Bidgoli, and A. Dahbashi, "Improving persian text classification using persian thesaurus", In Progress in Pattern Recognition, Image Analysis, Computer Vision,

زمینه‌های پژوهشی وی مباحثی نظیر الگوریتم‌های بهینه‌سازی پویا، رده‌بندی و خوشه‌بندی ترکیبی داده‌ها و پردازش سیگنال است.

نشانی رایانامه ایشان عبارت است از :

m.mohammadpour@iauyasooj.ac.ir

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

نشانی رایانامه ایشان عبارت است از:

parvin@iust.ac.ir

صمد نجاتیان تحصیلات خود را در مقطع کارشناسی در رشته مهندسی برق (الکترونیک) دانشگاه سیستان و بلوچستان در سال ۱۳۸۲ به پایان رساند. ایشان مدرک کارشناسی ارشد خود را در رشته برق (مخابرات)، از دانشگاه مشهد در سال ۱۳۸۶ اخذ کردند. وی مدرک دکترای تخصصی خود را در رشته برق (مخابرات)، در سال ۱۳۹۳ از دانشگاه UTM مالزی اخذ کردند. وی هم‌اکنون عضو هیئت علمی و معاونت پژوهشی دانشگاه آزاد اسلامی واحد یاسوج هستند.

نشانی رایانامه ایشان عبارت است از:

nejatian@iauyasooj.ac.ir

وحیده رضایی دارای مدرک تحصیلی در مقطع دکترای تخصصی رشته ریاضیات هستند. وی هم‌اکنون عضو هیئت علمی دانشگاه آزاد اسلامی واحد یاسوج است. زمینه‌های پژوهشی ایشان، بهینه‌سازی ریاضی، متن‌کاوی، پردازش سیگنال، داده‌کاوی و خوشه‌بندی داده‌ها

نشانی رایانامه ایشان عبارت است از:

v.rezaie@iauyasooj.ac.ir