یک روش نوین مستقل از دامنه کاربرد برای استخراج عبارات کلیدی از متون انگلیسی

حامد خوشدل
دانشگاه آزاد اسلامی واحد مشهد
hamed_khoshdel@yahoo.com

مجید وفایی جهان
دانشگاه آزاد اسلامی واحد مشهد
vafaeijahan@mshdiau.ac.ir

سید محمد حسین معطر
دانشگاه آزاد اسلامی واحد مشهد
moattar@mshdiau.ac.ir

چکیده:در این مقاله یک سیستم استخراج عبارات کلیدی معرفی می شود که به صورت مستقل از دامنه کاربرد، عبارات کلیدی را از یک متن ساده استخراج می‌کند. این متن ساده می‌تواند شامل روزنامه‌ها، وبلاگ‌ها، مقالات علمی و... باشد. روش پیشنهادی شامل ۳ مرحله است: در مرحله اول عبارات کاندید استخراج می‌شوند. در مرحله دوم برای هر کدام از عبارات کاندید، ویژگی‌های آماری (تعداد تکرار، اهمیت بخش‌ها، اولین رخداد و...) محاسبه می‌شوند و در مرحله سوم با امتیازدهی به عبارات کاندید بر اساس ویژگی‌های آماری محاسبه شده، از میان آن‌ها عبارات کلیدی استخراج می‌شوند. نوآوری و وجه تمایز روش پیشنهادی در به کار بردن ویژگی‌های آماری موثر شامل اهمیت بخش ها، آخرین رخداد و گستردگی می‌باشد. برای ارزیابی، روش پیشنهادی با ۲ نمونه از سیستم‌های استخراج عبارات کلیدی پرکاربرد موجود مقایسه می‌شود که نتایج حاکی از میزان بهبود ۱۴.۵ درصدی و ۲۳درصدی نسبت به این دو روش است.

واژه‌های کلیدی:عبارات کلیدی، استخراج عبارات کلیدی، متن کاوی، بدون نظارت، مستقل از دامنه کاربرد.

۱- مقدمه

یک عبارت کلیدی[۱]، عبارت کوتاهی است که معمولاً از یک، دو یا سه کلمه تشکیل شده است. عبارات کلیدی خلاصه کوتاه و مفیدی از متن مقاله را به محقق می‌دهند. عبارات کلیدی محتوای متن مقاله و عناوین بحث شده در آن را تا حدودی منعکس می‌کنند و محقق با بررسی و نگاه گذرا به آن‌ها می‌تواند تا حدودی به محتوای آن پی برده و تشخیص دهد که متن مقاله می‌تواند نیاز اطلاعاتی او را برآورده سازد یا نه. با این کار در وقت محققین صرفه جویی شده و آن‌ها در کم‌ترین زمان ممکن به مطلوب‌ترین اطلاعات خود دسترسی پیدا می‌کنند. از جمله کاربردهای عبارات کلیدی می‌توان به دسته بندی[۱]،خلاصه سازی [۲]، خوشه بندی [۳]، ایجاد لغت نامه [۴] و فرمول بندی مجدد پرس و جو در موتورهای جستجو به منظور بازگرداندن نتایج دقیق‌تر در مدت زمان کم‌تر [۵] اشاره کرد. هر وقت صحبت از عبارات کلیدی می‌شود اذهان عموم به سمت مقالات علمی روانه می‌شود در حالیکه عبارات کلیدی برای هر نوع متنی قابل استفاده هستند. این متون می‌توانند شامل مقالات علمی، مجلات، روزنامه‌ها، وبلاگ‌ها و... باشند. با وجود تمامی این دلایل مبنی بر ضرورت و اهمیت وجود عبارات کلیدی در متون، اما بیشتر نویسندگان تمایلی به استخراج عبارات کلیدی برای انواع متون نام‌برده ندارند. استخراج عبارات کلیدی به طور دستی کار بسیار زمان‌بر و مشکلی است بنابراین به سیستمی نیاز است که عبارات کلیدی را به طور خودکار از متون استخراج کند[۶].

از یک دیدگاه روش‌های استخراج عبارات کلیدی به دو دسته با نظارت[۲] و بدون نظارت[۳] تقسیم می‌شوند. در روش با نظارت یک مجموعه داده آموزشی وجود دارد که با یادگیری از آن‌ها، مدلی ساخته می‌شود و با به کار بردن این مدل بر روی سند جدید، عبارات آن به دو کلاس عبارات کلیدی و غیر کلیدی دسته بندی می‌شوند. کارایی این روش بسیار وابسته به داده‌های آموزشی است و عدم وجود داده‌های آموزشی باکیفیت منجر به افت کارایی سیستم استخراج عبارات کلیدی می‌شود. در این روش مدل ساخته شده مختص یک دامنه می‌باشد و به صورت وابسته به دامنه کاربرد عمل می‌کند. نمونه‌هایی [۶،۷] از این روش در بخش بعدی توضیح داده شده است. در روش بدون نظارت نیاز به داده‌های آموزشی نیست و با کمک استراتژی‌های رتبه دهی، مهم‌ترین عبارات از داخل متن استخراج می‌شوند. این روش بر خلاف روش با نظارت برای هر متنی از هر نوع دامنه‌ای، کاربرد دارد و به صورت مستقل از دامنه کاربرد عمل می‌کند. نمونه‌هایی [۸،۹،۱۰] از این روش نیز در بخش بعدی توضیح داده شده است.

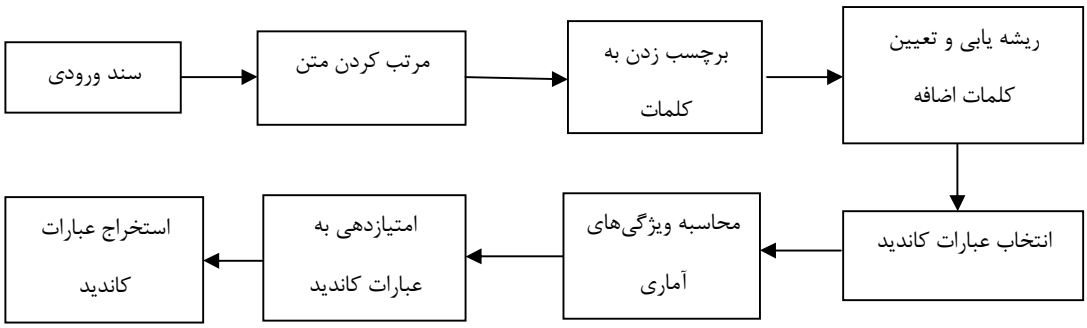
روش پیشنهادی در این مقاله برای استخراج عبارات کلیدی به صورت بدون نظارت بوده و بدون نیاز به داده‌های آموزشی به صورت مستقل از دامنه کاربرد عمل می‌کند. روش پیشنهادی شامل سه مرحله اصلی است: در مرحله اول عبارات کاندید استخراج می‌شوند. در مرحله دوم برای این عبارات کاندید ویژگی‌های آماری محاسبه شده و در مرحله سوم با امتیازدهی بر اساس ویژگی‌های آماری محاسبه شده و رتبه دهی به آن‌ها، از میان آن‌ها عبارات کلیدی استخراج می‌شوند. نوآوری و وجه تمایز روش پیشنهادی در به کار بردن ویژگی‌های آماری موثر مانند اهمیت بخش‌ها، آخرین رخداد و گستردگی می‌باشد. باقی‌مانده مقاله به این صورت است که در بخش ۲ کارهای پیشین بیان شده و در بخش ۳ به توصیف روش پیشنهادی با بیان جزئیات پرداخته می‌شود. در بخش ۴ ارزیابی و در بخش ۵ نتیجه گیری و ایده‌هایی برای توسعه روش پیشنهادی بیان می‌شود.

۲- کارهای پیشین

تورنی[۶] اولین کسی بود که استخراج عبارات کلیدی را به عنوان یک مسئله دسته بندی[۴] در نظر گرفت. او از ترکیب یکسری قوانین هیورستیک پارامتری و الگوریتم ژنتیک برای استخراج عبارات کلیدی استفاده کرد. KEA [۷] یکی از پر ارجاع‌ترین و ساده‌ترین روش‌های استخراج عبارات کلیدی است. در این الگوریتم از دو ویژگی TF * IDF و اولین رخداد استفاده می‌شود. با استفاده از داده‌های آموزشی و بر اساس تئوری بیزین یک دسته بندی کننده[۵]، ساخته شده و از آن برای استخراج عبارات کلیدی از سند جدید استفاده می‌شود. لیو در [۸] از تکنیک‌های خوشه بندی برای استخراج عبارات کلیدی استفاده کرد. روش او تضمین می‌کند این عبارات کلیدی، متن را از لحاظ معنایی پوشش می‌دهند. روش تارائو [۹] برای استخراج عبارات کلیدی از تکنیک‌های رتبه دهی مبتنی بر گراف استفاده می‌کند. در روش او هر سند به صورت گرافی از اصطلاحات و بر اساس ارتباط و وابستگی[۶](وزن) بین آن‌ها، نمایش داده می‌شود. سپس با استفاده از استراتژی‌های رتبه دهی و تخصیص امتیاز به هر نود، نودهایی با بالاترین رتبه به عنوان عبارات کلیدی شناخته می‌شوند. ارتباط (وزن) بین نودها بر اساس تعداد دفعات رخداد مشترک[۷]آن‌ها محاسبه می‌شود. براسول و همکارانش در [۱۰] ابتدا عبارات اسمی را از داخل سند استخراج کرد و سپس عباراتی که حداقل یک کلمه مشابه دارند را در داخل یک خوشه قرار داد. بر اساس تعداد تکرار عبارات اسمی و کلمات داخل آن، خوشه‌ها رتبه دهی می‌شوند و نهایتا n تا از خوشه‌ها با بالاترین رتبه به عنوان عبارات کلیدی انتخاب می‌شوند. وان [۱۱] یک ایده به روش تارائو [۹] اضافه کرد.او از تعدادی نزدیک‌ترین اسناد همسایه[۸] استفاده کرد تا با فراهم آوردن اطلاعات بیشتر، عبارات کلیدی بهتری استخراج شوند. در روش او هم از اطلاعات محلی سند اصلی و هم از اطلاعات سراسری اسناد همسایه در استخراج بهتر عبارات کلیدی استفاده می‌شود. در روشی مشابه [۱۲]وان، با اضافه کردن یک ایده به روش تارائو [۹] از تکنیک خوشه بندی برای پیدا کردن اسناد مشابه استفاده کرد. این اسناد مشابه اطلاعات بیشتری را برای استخراج بهتر عبارات کلیدی در اختیار قرار می‌دهند. تورنی در[۱۳] الگوریتم KEA [۷]را بهبود بخشید. ایده وی افزایش وابستگی میان عبارات کلیدی بود. روش او برای افزایش وابستگی، استفاده از پیوستگی آماری[۹] میان عبارات کاندید است تا عبارات کلیدی مستخرج، از لحاظ مفهومی به هم وابسته و نزدیک باشند. پیوستگی آماری با استفاده از وب کاوی ارزیابی می‌شود.

۳- روش پیشنهادی برای استخراج عبارات کلیدی

روش پیشنهادی بدون نظارت بوده و به تنهایی برروی یک متن ساده قابل اجراست یعنی بر خلاف روش‌های با نظارت، نیازی به مجموعه داده‌های آموزشی مناسب ندارد چرا که فراهم کردن داده‌های آموزشی مناسب کار زمان‌بر و مشکلی است و در صورتی که این داده‌ها کیفیت مناسب را نداشته باشند منجر به افت کارایی سیستم استخراج عبارات کلیدی با نظارت می‌شوند. کاربران می‌توانند روش پیشنهاد شده را بر روی انواع متون مانند روزنامه‌ها، وبلاگ‌ها و یا هر متن دلخواه دیگری استفاده کنند. مراحل روش پیشنهادی به طور مختصر در شکل ۱ نشان داده شده است. روش پیشنهادی شامل ۳ مرحله اصلی است: (۱) استخراج عبارات کاندید، (۲) محاسبه ویژگی‌های آماری برای هر عبارت کاندید، (۳) امتیازدهی به عبارات کاندید بر اساس ویژگی‌های آماری محاسبه شده و انتخاب تعدادی از آن‌ها با بالاترین رتبه به عنوان عبارات کلیدی. هر کدام از این مراحل سه‌گانه خودشان به چندین زیر مرحله تقسیم می‌شوند که به تفصیل توضیح داده می‌شوند.



شکل ۱)توصیف سیستم پیشنهادی که شامل سه مرحله اصلی است: انتخاب عبارات کاندید، محاسبه ویژگی‌های آماری، امتیازدهی و رتبه دهی

۳-۱- مرحله اول : استخراج عبارات کاندید

- **مرتب کردن[۱۰] متن:** در یک متن ممکن است بعضی از کلمات با حروف بزرگ آغاز شوند بنابراین برای داشتن متنی هماهنگ، تمامی کلمات متن با حروف کوچک نوشته می‌شوند.

- **برچسب زدن نحوی[۱۱]:** در یک متن هر کدام از کلمات نقشی (اسم، صفت، فعل و...) دارند که به کمک نرم افزار استنفورد [۱۴] نقش آن‌ها در متن تعیین می‌شود و به اصطلاح به هر کلمه برچسب زده می‌شود. در مراحل بعدی این برچسب‌ها در انتخاب عبارات کاندید مورد نظر استفاده می‌شوند.

- **ریشه یابی[۱۲] و تعیین کلمات اضافه[۱۳]:** فرض کنید در یک مقاله کلمه measure به شکل دیگری مثل measures به تعداد دفعات زیادی ظاهر می‌شود که باید این کلمه به ریشه خود یعنی measure تبدیل شود. برای ریشه یابی از الگوریتم پورتر [۱۵] استفاده می‌شود که این الگوریتم از یکسری قوانین گام به گام استفاده کرده و کلمه را به ریشه بر می‌گرداند. بعضی از کلمات مانند is, are, a, an, the, of و ... حروف اضافه هستند و هیچ تاثیری در انتخاب عبارات کلیدی ندارند. این کلمات تعیین می‌شوند. تعداد این کلمات در حدود ۴۲۵ تا می‌باشد که در [۱۶] به طور کامل ذکر شده است.

- **انتخاب عبارات کاندید:** بر اساس برچسب‌های زده شده توسط نرم افزار استنفورد [۱۴] عبارات کاندیدی استخراج می‌شوند که هر دو مورد زیر در آن‌ها صدق شود :

۱) تطبیق عبارات با یکی از حالت‌های N,NN,NNN,AN,ANN,.... که در این حالات N همان اسم و A همان صفت است. صفت می‌تواند شامل انواع صفت ساده، برتر و یا عالی باشد. اسم می‌تواند شامل انواع اسم ساده، جمع، خاص و... باشد. توصیف و تعداد کامل حالات در [۱۷] وجود دارد.

۲) کلمات اضافه در شروع یا پایان عبارات نباشند.

در مرحله انتخاب عبارات کاندید تمامی زیر توالی های موجود در یک عبارت کاندید استخراج می‌شوند.

۳-۲- مرحله دوم : محاسبه ویژگی‌های آماری

با انتخاب عبارات کاندید در مرحله قبل، در این مرحله برای هر کدام از آن‌ها ویژگی‌های آماری محاسبه می‌شوند. این ۵ ویژگی عبارتند از: تعداد تکرار، اهمیت بخش‌ها (وزن دهی به بخش‌ها)، اولین رخداد، آخرین رخداد و طول عمر[۱۴](گستردگی[۱۵]).

- **تعداد تکرار × وزن بخش‌ها**

وزن بخش‌ها: احتمال قرار گرفتن عبارات کلیدی در بخش‌های متفاوت یک متن، با توجه به میزان اهمیت آن‌ها متفاوت است و بر همین اساس، بخش‌ها، وزن‌های مختلفی به خود می‌گیرند. واضح است که بخش‌هایی مثل عنوان، چکیده، مقدمه و نتیجه گیری از اهمیت بالایی برخوردار هستند و احتمال حضور عبارات کلیدی در آن‌ها زیاد است پس وزن بالایی به خود می‌گیرند.

تعداد تکرار: تعداد تکرار هر عبارت کاندید در هر بخش از متن به صورت مجزا، در نظر گرفته می‌شود.

با در نظر گرفتن دو معیار فوق ویژگی اول برای هر کدام از عبارات کاندید بر اساس فرمول زیر محاسبه می‌شود:

$$F_1 = \frac{\sum_{i=1}^{\text{تعداد بخش ها}} \text{وزن بخش} \times \text{تعداد تکرار}}{\text{مقدار ماکزیمم}} \quad (۱)$$

دلیل قرار گرفتن مقدار ماکزیمم در مخرج فرمول بالا این است که F_1 عددی نرمال در بازه $[0,1]$ باشد.

- **اولین رخداد:** در یک متن یک عبارت کاندید هر اندازه مهم‌تر و کلیدی‌تر باشد زودتر پدیدار شده و نویسنده زودتر آن را بیان می‌کند. بنابراین انتظار می‌رود عبارات کلیدی و مهم در بخش‌های ابتدایی متن مثل چکیده و مقدمه پدیدار شوند. این ویژگی برای عبارات کاندید به صورت زیر محاسبه می‌شود:

$$F_2 = 1 - \frac{\text{تعداد کلمات جلوتر از عبارت کاندید}}{\text{کل کلمات متن}} \quad (۲)$$

مقدار این ویژگی هر اندازه بیشتر باشد یعنی عبارت کاندید در متن زودتر پدیدار می‌شود. این ویژگی مقداری نرمال شده در $[0,1]$ است.

- **آخرین رخداد:** یک عبارت کاندید هراندازه مهم‌تر و کلیدی‌تر باشد تا انتهای متن نقش خودش را ایفا کرده و پدیدار می‌شود. بنابراین انتظار می‌رود عبارات کلیدی در بخش‌های پایانی متن مثل نتیجه گیری و کارهای آینده پدیدار شوند. این ویژگی برای عبارات کاندید به صورت زیر محاسبه می‌شود:

$$F_3 = \frac{\text{تعداد کلمات جلوتر از عبارت کاندید}}{\text{کل کلمات متن}} \quad (۳)$$

این ویژگی نیز مقداری نرمال شده در $[0,1]$ است.

- **طول عمر(گستردگی):** یک عبارت کاندید هراندازه مهم‌تر و کلیدی‌تر باشد در یک متن به مدت طولانی‌تری نقش ایفا کرده و زندگی می‌کند و به بیان بهتر میزان بیشتری از متن را پوشش می‌دهد. این ویژگی برای عبارات کاندید به صورت زیر محاسبه می‌شود:

$$F_4 = \frac{\text{تعداد کلمات بین اولین و آخرین رخداد عبارت کاندید}}{\text{کل کلمات متن}} \quad (۴)$$

هر چقدر مقدار این ویژگی بالاتر باشد، عبارت کاندید مقدار بیشتری از متن را پوشش می‌دهد. این ویژگی نیز مقداری نرمال شده در $[0,1]$است.

۳-۳- مرحله سوم : امتیازدهی به عبارات کاندید

با استفاده از ویژگی‌های آماری محاسبه شده برای هر عبارت کاندید، آن‌ها امتیازدهی می‌شوند. در مرحله بعد، از این امتیازدهی برای استخراج مناسب‌ترین عبارات به عنوان عبارات کلیدی استفاده می‌شوند.

در حالت اول هر ۴ ویژگی آماری ذکر شده اهمیت یکسان دارند در این صورت امتیاز هر عبارت کاندید به صورت زیر محاسبه می‌شود:

$$\text{score(phrase)} = \frac{\sum_{i=1}^{4} F_i}{4} \quad (۵)$$

در حالت دوم هر کدام از ۴ ویژگی آماری به نسبت اهمیت و تأثیرگذاری در انتخاب عبارات کلیدی، یک وزنی به آن‌ها اختصاص داده می‌شود. در این صورت امتیاز هر عبارت کاندید به صورت زیر محاسبه می‌شود:

$$\text{score(phrase)} = \frac{\sum_{i=1}^{4} F_i * W_i}{\sum_{i=1}^{4} W_i} \quad (۶)$$

در هر دو فرمول بالا F_i مقدار ویژگی آماری و در فرمول بالا W_i وزن متناظر با هر ویژگی آماری است.

- استخراج عبارات کلیدی: برای انتخاب n تا از عبارات کاندید به عنوان عبارات کلیدی، ابتدا بر اساس امتیاز مرتب شده به صورت نزولی، عبارات کاندید مرتب و رتبه دهی می‌شوند و سپس از استراتژی مطرح شده در [۱۸] استفاده می‌شود. ۰.۴n عدد از عبارات کاندید یک کلمه‌ای با بالاترین امتیاز، ۰.۴n عدد از عبارات کاندید دو کلمه‌ای با بالاترین امتیاز و ۰.۲n عدد از عبارات کاندید سه کلمه‌ای با بالاترین امتیاز به عنوان عبارات کلیدی نهایی انتخاب می‌شوند.

۴- ارزیابی

۴- ۱- مجموعه داده

مجموعه داده مورد استفاده برای ارزیابی کارایی و عملکرد روش پیشنهادی برای استخراج عبارات کلیدی، ۲۱۵ عدد مقاله است که در مورد حوزه‌های مختلف علم کامپیوتر می‌باشد[۱۹]. مجموعه داده از کیفیت و کارایی بالایی برخوردار بوده و توسط [۲۰] مورد استفاده قرار گرفته است. هر مقاله در مجموعه داده شامل دو گروه عبارات کلیدی می‌باشد: گروه اول که توسط نویسنده مقاله تخصیص داده می‌شود و گروه دوم توسط افرادی تخصیص داده می‌شود که این افراد کاملاً خبره و آشنا به حوزه‌های مختلف علم کامپیوتر هستند. دلیل استفاده از این مجموع عبارات کلیدی تخصیص یافته توسط افراد خبره این است که هر چند طبق اظهارات پاینتر [۲۱]، عبارات کلیدی تخصیص یافته توسط نویسنده مقاله، نمایندگان خوبی برای معرفی موضوع و محتوای مقاله هستند اما ممکن است نتوانند تمامی عبارات کلیدی مفید را پوشش کامل دهند چون هر نویسنده از دیدگاه شخصی خودش عبارات کلیدی را استخراج می‌کند[۲۰]. روش پیشنهادی با شمارش تعداد تطبیق‌های بین عبارات کلیدی استخراج شده توسط روش پیشنهادی و مجموع دو گروه عبارات کلیدی تخصیص یافته به هر مقاله، ارزیابی می‌شود.

۴– ۲- نتایج

برای ارزیابی کارایی و عملکرد روش پیشنهادی دو آزمایش انجام شده است. در آزمایش اول، روش‌های استخراج عبارات کلیدی در مقالات KEA[۷] وNguyen&Kan[۲۰] به عنوان روش‌های پایه‌ای برای ارزیابی و مقایسه با روش پیشنهادی، در نظر گرفته می‌شوند. در این آزمایشات از مجموعه داده ذکر شده ۱۲۰ مقاله انتخاب می‌شود. جدول ۱ تعداد میانگین تطبیق‌های واقعی بین ۳ الگوریتم را نشان می‌دهد به شکلی که برای هر مقاله موجود در مجموعه داده ۱۰ عبارت کاندید اول با بالاترین رتبه به عنوان عبارات کلیدی استخراج می‌شوند. با توجه به جدول ۱ برای روش پیشنهادی، میزان بهبود ۱۴.۵درصدی نسبت به روش Nguyen&Kan [۲۰] و میزان بهبود ۲۳ درصدی نسبت به روش KEA[۷] دیده می‌شود.

در آزمایش دوم تنها روش KEA[۷] مورد مقایسه قرار گرفته و کل ۲۱۵ مقاله موجود در مجموعه داده استفاده می‌شوند که از میان آن‌ها، ۷۰ مقاله برای یادگیری الگوریتم KEA[۷] انتخاب می‌شوند. برای هر مقاله ۷، ۱۵ و ۲۰ عبارت کاندید اول با بالاترین رتبه به عنوان عبارات کلیدی استخراج می‌شوند. نتایج این آزمایش در جدول ۲ قرار داده شده است. واضح است که دلیل برتری عملکرد روش پیشنهادی نسبت به KEA، مستقل بودن آن از دامنه کاربرد و عدم نیاز به فعالیت‌های آموزشی است. قابل ذکر است که روش KEA یکی از پر ارجاع‌ترین و پر مقایسه‌ترین روش‌ها در حوزه استخراج عبارات کلیدی است[۵].

جدول۱)عملکرد روش پیشنهادی در مقایسه با دو روش دیگر

میانگین تعداد تطبیق‌های واقعی	الگوریتم
۳.۰۳	KEA
۳.۲۵	Nguyen&Kan
۳.۷۳	روش پیشنهادی

جدول۲)عملکرد روش پیشنهادی در مقایسه با KEA

تعداد عبارات کلیدی استخراج شده	میانگین تعداد تطبیق‌های واقعی	
	KEA	روش پیشنهادی
۷	۲.۰۵	۲.۵۳
۱۵	۲.۹۵	۳.۹۴
۲۰	۳.۰۸	۴.۰۳

یک نمونه خروجی از روش پیشنهادی برای استخراج عبارات کلیدی در جدول ۳ نشان داده شده است. روش پیشنهادی برای ۳ مقاله موجود در مجموعه داده مذکور اجرا شده و برای هر کدام ۱۰ عبارت کاندید اول با بالاترین رتبه استخراج می‌شوند. در جدول ۳ ردیف اول عنوان مقالات را مشخص می‌کند. ردیف دوم و سوم مجموع عبارات کلیدی تخصیص یافته به هر مقاله توسط دوگروه نویسنده و افراد خبره می‌باشد. ردیف چهارم مجموع ۱۰ عبارت کلیدی استخراج شده توسط روش پیشنهادی می‌باشد و عبارات پررنگ همان عبارات تطبیقی هستند.

جدول۳)نمایش نمونه خروجی از روش پیشنهادی

عنوان و شماره مقاله	Development of E-commerce Statistics and the Implications #66	A Geometric Constraint Library for 3D Graphical Applications #12	Information Revelation and Privacy in Online Social Networks 113#
عبارات کلیدی تخصیص یافته توسط نویسنده	e-commerce, e-commerce statistics, statistical survey.	geometric constraints, constraint satisfaction, geometric layout, 3D graphics, scene graphs.	Online privacy , information revelation, social networking Sites.
عبارات کلیدی تخصیص یافته توسط افراد خبره	China, Statistics Survey ,E-commerce ,Implications Development stage, Authority, Definition of E-commerce, Statistical methods, Measurement.	3D graphical applications, geometric constraints, layout, behaviors, graphical objects, coordinate transformation.	privacy information revelation, online behavior, college, privacy risk, online social networking, social network theory, data visibility, privacy preference, stalking, re-identification, face book.
عبارات کلیدی استخراج شده توسط روش پیشنهادی	commerce **,e-commerce, implication,** country **, ecommerce statistics** , developed country, **statistical survey,** perfect ecommerce, annual statistical survey , annual sample investigation	Constraint, object, chorus, chorus3d **, geometric constraint, graphical object,** graphical application **, constraint satisfaction, 3D graphical application ,** geometric constraint library**.**	Network, form**, privacy ,**information, social network **, information revelation,** privacy implication **, privacy preference,** online social network **, social network theory.**

۵- نتیجه گیری و کارهای آینده

در این مقاله یک روش بدون نظارت و مستقل از دامنه کاربرد برای استخراج عبارات کلیدی ارائه شدکه بر خلاف روش‌های با نظارت نیازی به داده‌های آموزشی ندارد و برای یک متن ساده قابل اجراست. به طور خلاصه روش پیشنهادی شامل ۳ مرحله است: در مرحله اول عبارات کاندید استخراج می‌شوند. در مرحله دوم برای هر کدام از عبارات کاندید ویژگی‌های آماری محاسبه شده و در مرحله سوم با امتیازدهی به آن‌ها، از میان آن‌ها عبارات با بالاترین رتبه به عنوان عبارات کلیدی استخراج می‌شوند. نوآوری و وجه تمایز روش پیشنهادی در به کار بردن ویژگی‌های آماری موثر مانند اهمیت بخش‌ها، آخرین رخداد و گستردگی می‌باشد. روش مذکور بر روی ۱۲۰ مقاله علمی اجراشده و با ۲ تا از سیستم‌های استخراج عبارات کلیدی پرکاربرد موجود مقایسه می‌شود که نتایج حاکی از میزان بهبود ۱۴.۵ درصدی و ۲۳ درصدی نسبت به این دو روش است.

همان طور که در ارزیابی اشاره شد عبارات کلیدی تخصیص یافته توسط نویسنده نمی‌توانند تمامی عبارات کلیدی مفید را به طور کامل پوشش دهند بنابراین باید به دنبال تکنیک و روشی بود که تعداد بهینه عبارات کلیدی را برای هر متنی تعیین کند تا مسئله پوشش کامل عبارات کلیدی مفید ارضا شود و این می‌تواند به عنوان اولین ایده برای توسعه روش پیشنهادی باشد. ایده دوم این است که باید با استفاده از تکنیک و الگوریتمی مشخص، میزان تأثیرگذاری هر کدام از ویژگی‌های آماری در استخراج عبارات کلیدی تعیین شده و به آن‌ها وزنی مناسب اختصاص داده شود.

مراجع

[1] Krulwich, B., Burkey, C., "Learning user information interests through the extraction of semantically significant phrases" In: Hearst, M., Hirsh, H. (eds.) AAAI 1996 Spring Symposium on Machine Learning in Information Access, pp. 110–112. AAAI Press, California (1996)

[2] Berger, A.L., Mittal, V.O., Ocelot, "A system for summarizing web pages". Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval,. ACM, New York ,pp. 144–151,2000.

[3] Hammouda, K.M., Matute, D.N., Kamel, M.S, "Corephrase: Keyphrase extraction for document clustering", Perner, P., Imiya, A. (eds.) MLDM 2005. LNCS (LNAI), Springer, Heidelberg,vol. 3587, pp. 265–274.,2005.

[4] Kosovac, B., Vanier, D.J., Froese, "T.M.: Use of keyphrase extraction software for creation of an AEC/FM thesaurus". Electronic Journal of Information Technology in Construction 5, pp. 25–36, 2000.

[5] Lui Y., Calinescu A., Brent R. and Yang L., "Extraction of Significant Phrases from Documents in English and Chinese", Cybernetics and Systems,2008.

[6] Turney, P.D.: "Learning algorithms for keyphrase extraction", Information Retrieval, 2(4), pp.303–336,2000.

[7] Witten, I.H., Paynter, G.W., Frank, E., Gutwin, C., Nevill-Manning, C.G., "Kea: practical automatic keyphrase extraction", Proceedings of the fourth ACM conference on Digital libraries, ACM, New York, pp. 244–255, 1999.

[8] Liu, Z., Li, P., Zheng, Y., Sun, M., "Clustering to find exemplar terms for keyphrase extraction", Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, ACL, Singapore, pp. 257–266, 2009.

[9]Mihalcea R., Tarau P., "Text rank: Bringing order into texts". In Proceedings ofthe *2004* Conference on Empirical Methods in Natural Language Processing, 2004.

[10] Bracewell, D.B., Ren, F., Kuroiwa, S., "Multilingual single document keyword extraction for information retrieval", Proceedings of the 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, Wuhan, pp. 517–522, 2005.

[11] Xiaojun Wan, Jianguo Xiao, "Single document keyphrase extraction using neighborhood knowledge", Proceedings of the Twenty-ThirdAAAI Conference on Artificial Intelligence, pp.855–860,2008.

[12] Xiaojun Wan,Jianguo Xiao. "Collabrank: Towards a collaborative approach to single document keyphrase extraction", Proceedings of COLING, pp.969–976,2008.

[13]Turney,P.D., "Coherence keyphrase extraction via web mining", IJCAI 2003, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence,pp.434-439, 2003.

[14]Stanford POS tagger, http: //nlp.stanford.edu/software/tagger .

[15] Porter, M.F., "An algorithm for suffix stripping", Readings in information retrieval, pp.313–316, 1997.

[16] FRANCIS, W., KUCERAH., Frequency Analysis of English Usage, New York, Houghton Mifflin,1982.

[17] Justeson, J., Katz, S., "Technical terminology: some linguistic properties and an algorithm for identification in text", Natural Language Engineering 1, pp.9–27,1995.

[18] Kumar, N., Srinathan, K., "Automatic keyphrase extraction from scientific documents using n-gram filtration technique", Proceedings of the Eight ACM symposium on Document engineering,ACM, New York, pp. 199–208, 2008.

[19] keyphrase Corups, http: wing.comp.nus.edu.sg/downloads.
[20] Nguyen, T.D., Kan, M.Y. ",Keyphrase extraction in scientific publications", Goh, D.H.L., Cao, T.H., Sølvberg, I., Rasmussen, E.M. (eds.) ICADL 2007, LNCS, Springer, Heidelberg,vol. 4822, pp. 317–326, 2007.
[21] Jones, S., Paynter, G.W., "Human evaluation of Kea, an automatic keyphrasing system", ACM/IEEE Joint Conference on Digital Libraries, pp. 148–156, 2001.

.

زیر نویس:

[۱] Key phrase
[۲] Supervised
[۳] Unsupervised
[۴] Classification
[۵] Classifier
[۶] Coherence
[۷] Co occurrence
[۸] Nearest neighbor
[۹] Statistical association
[۱۰] Case fold
[۱۱] Pos tagging
[۱۲] Stemming
[۱۳] Stop word
[۱۴] Life time
[۱۵] Spread