مراکز آموزش عالی ایران. *مدیریت و برنامه ریزی در نظام های آموزشی*، ۱۱(۲)، ۷۶-۵۱.

Gadagin, B. R. & Selvaraj, C. (2016). A Comparative Study of Plagiarism Detective Software (PDS) Tools and Techniques. *International research journal of multidisciplinary studies*, 2(3), *1-5.*

Hu, G. (2015). Research on plagiarism in second language writing: Where to from here? *Journal of Second Language Writing*, 30, 100-102.

Patil1, A. & Bomanwar, N. (2016). Survey on Different Plagiarism Detection Tools and Software's, International. *Journal of Computer Science and Information Technologies*, 7 (5), 2191-2193.

Shkodkina, Y. & Pacauskas, D. (2017). Comparative Analysis of Plagiarism Detection Systems *Business Ethics and Leadership*, 1(3), 27-3.

معرفی مدلی زبان‌شناسی- آماری برای استخراج خودکار کلیدواژه‌ها

الهام علایی ابوذر[۱]

چکیده

کلیدواژه‌ها نقش مهمی در استخراج اطلاعات درست از متون ایفا می‌کنند. هر روز کتاب‌ها و مقالات بسیاری منتشر می‌شوند، خواندن همهٔ متن کتاب یا مقاله برای کسب اطلاعاتی که در جستجوی آن هستیم، کاری بسیار دشوار است؛ استفاده از روش‌هایی که بتوان کلمات کلیدی یک متن را استخراج کرد یا خلاصه‌ای از متن ارائه داد، گامی بزرگ در جهت دسترسی به اطلاعات مورد جستجو است. در پژوهش حاضر مدلی مبتنی بر دانش زبان‌شناسی (مقولهٔ اجزای واژگانی کلام) و آمار ارائه می‌گردد. روش پیشنهادی را می‌توان با استفاده از نرم‌افزارهای مربوطه انجام داد و سرعت استخراج کلیدواژه‌ها را از متون بالا برد.

واژه های کلیدی: کلمات کلیدی، مدل مبتنی بر دانش زبانی، مقولهٔ اجزای واژگانی کلام، آمار، استخراج کلیدواژه‌ها.

۱. استادیار پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، Alayi@irandoc.ac.ir

۱- مقدمه

استخراج کلیدواژه[1] به معنی استخراج آن دسته از کلمات یک متن است که بتواند شاخصی برای محتوای آن متن در نظر گرفته شود. در واقع، استخراج کلیدواژه، روشی است برای تشخیص خودکار مجموعه‌ای از عبارات که بهترین توصیف را از موضوع یک سند ارائه دهند. از آنجائیکه کلیدواژه‌ها اغلب عبارات دو یا چندکلمه‌ای هستند، می‌توان آن‌ها را «عبارات کلیدی[2]» نیز نامید (ترنی: ۲۰۰۰). عبارت کلیدواژه، مترادف‌هایی دارد که شامل: عبارات کلیدی، بخش‌های کلیدی[3]، الفاظ کلیدی[4]، یا کلمات کلیدی[5] است؛ همهٔ این موارد مترادف، یک کارکرد واحد دارند: مشخص کردن موضوع مورد بحث در متن/سند (بلیگا: ۲۰۱۴). جستجوی اطلاعات در اینترنت با استفاده از کلیدواژه‌ها صورت می‌پذیرد؛ در واقع کلیدواژه‌ها، خلاصه‌ای بسیار فشرده از یک متن ارائه می‌دهند و جهت بازیابی بهتر و سریع‌تر اطلاعات مورد استفاده قرار می‌گیرند. از آنجائیکه تعداد معدودی از اسناد/ متون حاوی کلیدواژه هستند، طراحی روشی که بتوان به طور خودکار کلیدواژه‌های یک متن را استخراج کرد، کمک بزرگی به بهبود بازیابی اطلاعات در محیط الکترونیکی می‌کند (هولث: ۲۰۰۳). استخراج کلیدواژه‌ها اهداف خاصی را دنبال می‌کند که از میان آن‌ها می‌توان به موارد زیر اشاره کرد (ترنی: ۲۰۰۰):

- زمانی‌که روی صفحهٔ اول یک مقالهٔ چاپ شده (پس از چکیده)، کلیدواژه‌ها ظاهر می‌شوند، هدف، خلاصه‌سازی است؛ خوانندهٔ مقاله به کمک این کلیدواژه‌ها می‌تواند به سهولت و به سرعت دریابد آیا مقالهٔ حاضر در حوزهٔ علاقهٔ وی گنجانده می‌شود یا خیر.
- زمانی‌که کلیدواژه‌ها در فهرست یک مجلهٔ چاپ شده قرار می‌گیرند، هدف، نمایه‌سازی است؛ خواننده‌ای که در جستجوی اطلاعات خاص است، به سرعت می‌تواند مقالهٔ مربوط به حوزهٔ مورد جستجو را پیدا کند.
- زمانی‌که یک موتور جستجو مجهز به بخشی تحت عنوان «کلیدواژه‌ها» باشد، هدف آن، کمک به خواننده برای جستجوی دقیق‌تر و جامع‌تر اطلاعات است. به عنوان مثال، می‌توان از این ویژگی در مرورگرهای وب استفاده کرد؛ بدین ترتیب که کاربر با فشار دادن یک دکمه، عبارت‌های کلیدی متن را مشاهده و در نتیجه به حوزهٔ موضوعی متن مورد نظر پی می برد.

---

1 Key word extraction
2 Key phrases
3 Key segments
4 Key terms
5 Key words

حضور عبارت‌های کلیدی در نتایج جستجو می‌تواند به اصلاح و تعریف مجدد فرمول جستجو و حتی تغییر دیدگاه کاربران از ساختار موجود در یک زمینهٔ خاص کمک کند؛ یعنی کاربران می‌توانند با افزودن یا حذف واژگان، دامنهٔ جستجو را محدودتر کرده، ضریب دقت را بالاتر ببرند. بنابراین می‌توان عبارت‌های کلیدی را به عنوان جزئی لازم برای سیستم‌های بازیابی اطلاعات معرفی کرد (گزنی، ۱۳۸۵). هدف کلی استخراج کلیدواژه‌ها، یافتن مجموعه‌ای از کلمات است که بهترین توصیف از یک متن خاص، صرفنظر از حوزهٔ تخصصی متن، ارائه دهد.

از آنجائیکه کلیدواژه‌ها کاربردهای فراوانی در به کارگیری مستندات الکترونیکی دارند، شناسایی روش‌های خودکار و بهبودیافته برای استخراج این دسته از کلمات همیشه مورد توجه بوده است. هر چه این امر به سمت خودکار شدن پیش رود، باعث سرعت بخشیدن به پردازش‌های گوناگون روی متون و بازیابی اطلاعات می‌شود و کاربران می‌توانند با سرعت و دقت بالاتری اطلاعات مربوطه را از متون گوناگون الکترونیکی استخراج کنند. در زبان فارسی کلمات دارای صورت‌های نگارشی متنوعی هستند و پوشش کلیه حالات دستوری کلمات با به کارگیری تعدادی قواعد معین ناممکن است، به همین دلیل استخراج کلمات کلیدی به طور خودکار از متون فارسی دشوار و پیچیده است. برای استخراج کلیدواژه‌های متون فارسی، پژوهش‌هایی انجام شده است که به برخی از آن‌ها اشاره می‌شود:

گزنی (۱۳۸۵) از روش‌های آماری، نحوهٔ توزیع واژگان، مجاورت و ... برای استخراج عبارات کلیدی از مقاله‌های فارسی استفاده می‌کند. سیستمی که بر پایه پژوهش وی طراحی گردیده، با توجه به بازخوردهای کاربر از قابلیت یادگیری برخوردار است؛ به گونه‌ای که در طول زمان مرتباً به کارایی آن افزوده می‌شود. اسماعیلی آدرگانی و فراهی (۱۳۹۲) معتقدند توجه به معنای جملات و گلوگاه‌های معنایی، مانند بخش‌های نتیجه‌گیری و بخش‌هایی که مؤلف در تلاش برای اثبات ادعا و ارائه دلایل است، می‌تواند سرنخ‌های خوبی را برای رسیدن به کلیدواژه‌ها ارائه کند. بخش‌هایی از یک مقاله که حاوی کنش‌های گفتار، عناصر استدلال و نقل قول‌ها هستند، جزء پرمعناترین بخش‌های یک مقاله محسوب می‌شوند؛ در این بخش‌ها کلمات اصلی و مرتبط با موضوع مقاله از فراوانی بالایی برخوردار هستند. همچنین افعال پرتکرار و لحاظ کردن قطبیت آن‌ها و توجه به جایگاه جملات در متن به عنوان راه‌های بعدی در رسیدن به کلیدی‌ترین کلمات محسوب می‌شوند. معادی و فولادی قلعه (۱۳۹۴) روشی جدید برای استخراج کلمات کلیدی با استفاده از ویژگی‌های آماری و بردار رخداد کلمه در هر متن، ارائه می‌دهند.

این روش برای زبان فارسی بر روی متن منفرد و بدون درنظر گرفتن دامنهٔ موضوعی متون اجرا می‌شود. این پیاده‌سازی با مجموعه داده‌های تشکیل شده برای این پژوهش که دربرگیرندهٔ ۱۰۰ مقالهٔ معتبر فارسی است، ارزیابی و با کلمات کلیدی مشخص شده توسط نویسنده هر مقاله مقایسه شده است و معیارهای بازیابی و دقت محاسبه شده برای کل مجموعه داده نتایج قابل توجهی را نشان می‌دهد. راد و همکاران (۱۳۹۵) سعی دارند با استفاده از اطلاعات زبان-شناختی و اصطلاح‌نامه، روش جدیدی برای استخراج کلمات کلیدی بامعناتر ارائه دهند. با استفاده از اصطلاح‌نامه که از نظامی ساختارمند برخوردار است، می‌توان شبکهٔ کلمات کلیدی، شامل کلمات هم‌ارز، کلمات سلسله‌مراتبی و وابسته را تکمیل کرده و افزایش داد. بنابراین می-توان توافق بین جستجوی کاربران و کلمات کلیدی متنی را بیشتر کرد و جامعیت جستجو را افزایش داد.آن‌ها در روش پیشنهادی خود اینگونه عمل می‌کنند: در گام نخست کلمات غیرمهم و عمومی حذف می‌شوند، سپس کلمات متن ریشه‌یابی و در ادامه برای مشخص شدن اهمیت نسبی کلمات با استفاده از روش‌های وزن‌دهی، یک وزن عددی به هر کلمه منسوب می‌شود که بیان‌گر میزان تاثیر کلمه در ارتباط با موضوع متن و در مقایسه با سایر کلمات به کار رفته در متن است. مجموعه عملیات مذکور، به خصوص استفاده از اصطلاح‌نامه، باعث می‌شود دسته-بندی متون دقیق‌تر انجام گیرد و به نوعی ردهٔ علمی سلسله‌مراتبی متون در حوزهٔ بازیابی اطلاعات نیز مشخص می‌شود. نتایج آزمایش‌ها روی چندین متن در موضوعات مختلف نشان دهندهٔ دقت و توانایی روش پیشنهادی در استخراج کلمات کلیدی منطبق با خواست کاربر است و در نتیجه خوشه‌بندی دقیق‌تر متون انجام می‌شود.

در پژوهش حاضر روشی مبتنی بر اطلاعات زبان‌شناسی و آمار برای استخراج خودکار کلیدواژه‌ها در متون فارسی پیشنهاد می‌شود.

۲- روش پژوهش و معرفی مدل پیشنهادی برای استخراج خودکار کلیدواژه‌ها

به طور کل، رویکردهای متفاوتی برای استخراج خودکار کلیدواژه‌ها پیشنهاد شده است که می‌توانند در چهار دسته قرار گیرند (صدیقی و شاران، ۲۰۱۵):

۱- رویکردهای قاعده-بنیاد مبتنی بر دانش زبان‌شناسی[1]: این رویکردها عموماً قاعده-بنیاد هستند و از دانش/ مولفه‌های زبان‌شناسی بهره می‌برند. در این رویکردها، از مشخصه‌های

1 Rule-based linguistic approaches

زبان‌شناسی کلمات (نوعاً در بافت نحوی) و از تحلیل‌های نحوی، واژگانی، گفتمانی و ....استفاده می‌شود (کر و گوپتا، ۲۰۱۰). این روش‌ها ممکن است دقیق‌تر باشند، اما مستلزم استفاده از دانش حوزهٔ خاص و دانش زبان‌شناسی هستند. در واقع این روش‌ها، از مولفه‌های زبانی موجود در کلمات و جملات برای استخراج کلیدواژه‌ها استفاده می‌کنند.

۲- رویکردهای آماری[1]: این رویکردها عموماً مبتنی بر پیکره‌های زبانی و مولفه‌های آماری استخراج شده از پیکره هستند. مهم‌ترین مزیت این روش‌ها این است که مستقل از زبانی هستند که این رویکردها در موردشان به کار گرفته می‌شود؛ بنابراین تکنیک‌های یکسان دربارهٔ زبان‌های گوناگون به کار گرفته می‌شود. نتیجهٔ به کار گیری این روش‌ها، در مقایسه با روش-های زبان‌شناسی، ممکن است درست و دقیق نباشد، اما در دسترس بودن مجموعه‌های بزرگ از داده‌ها، این امکان را فراهم می‌آورد که بتوان تجزیه و تحلیل آماری انجام داد و به نتایج مطلوبی رسید. در روش آماری، نیازی به داده‌های آزمون نیست؛ اطلاعات آماری کلمات برای تشخیص کلیدواژه‌های یک متن، مورد استفاده قرار می‌گیرد؛ از روش‌های معمول این رویکرد می‌توان به N-Gram، فراوانی کلمات، باهم‌آیی کلمات و....اشاره کرد.

۳- رویکردهای یادگیری ماشینی[2]: این رویکردها عموماً روش‌های یادگیری باسرپرست/ بانظارت[3] را به کار می‌گیرند که در این روش‌ها، کلیدواژه‌ها از سندها/متونی که برای آموزش ماشین به کار گرفته می‌شوند، استخراج می‌شوند تا مدل اولیه به دست آید، در نهایت مدل با دادهٔ آزمون سنجیده می‌شود و پس از ساخت مدل، برای استخراج کلیدواژه‌ها از متون جدید به کار گرفته می‌شود. این رویکردها شامل مدل بیزی ساده[4]، ماشین بردار حمایت[5] و....است. این رویکردها مستلزم استفاده از پیکره‌های برچسب‌خورده است که ساخت چنین پیکره‌هایی دشوار است. در صورت فقدان چنین پیکره‌هایی از روش‌های یادگیری بدون سرپرست/ بدون نظارت[6]، یا نیمه‌نظارتی/ نیمه‌سرپرستی[7] استفاده می‌شود.

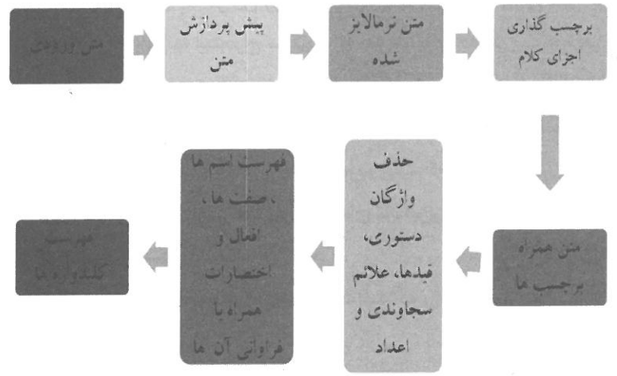
۴- رویکردهای حوزه-محور[8]: رویکردهایی هستند که با استفاده از دانش پیش‌زمینهٔ یک

1 Statistical approaches
2 Machine learning approaches
3 Supervised learning
4 Naïve Bayes
5 Support vector machine
6 Unsupervised learning
7 Semi-supervised learning
8 Domain specific approaches

حوزهٔ خاص (در قالب هستی‌شناسی و...) عمل می‌کنند.

معروف‌ترین این روش‌های، رویکرد یادگیری ماشینی است. استخراج خودکار کلیدواژه‌ها با استفاده از یادگیری ماشینی اولین بار توسط ترنی (۲۰۰۰) معرفی شد. همانگونه که ذکر شد، در این روش عمل یادگیری توسط آموزش ماشین با استفاده از اسنادی که کلیدواژه‌هایشان مشخص شده است، صورت می‌پذیرد. سپس، مدل آموزش‌دیده روی اسنادی که کلیدواژه‌هایشان مشخص نیست، به کار گرفته می‌شود که در نهایت کلمات داخل متن به صورت کلیدواژه یا غیرکلیدواژه برچسب‌گذاری می‌شوند؛ اگر از مدل احتمالاتی استفاده شده باشد، کنار هر کلمه، احتمال اینکه آن کلمه کلیدواژه باشد، درج می‌شود (هولت: ۲۰۰۳). در اکثر روش‌های معروف، تعداد کلمات استخراج شده به عنوان کلیدواژه ۱۰ الی ۱۵ کلمه است. بیشتر روش‌های استخراج کلیدواژه‌ها مبتنی بر پردازش زبان طبیعی، از فرهنگ‌های لغت برای مشخص کردن ریشهٔ کلمات و بخش‌های گفتار استفاده می‌کنند.

روش پیشنهادی در پژوهش حاضر برای استخراج کلیدواژه‌ها از متون فارسی، مبتنی بر دانش زبان‌شناسی و اطلاعات آماری (نوعاً فراوانی کلمات) است؛ این مدل پیشنهادی در شکل زیر قابل مشاهده است:



برای استخراج کلیدواژه‌ها ابتدا باید پیش‌پردازش‌هایی روی متن انجام شود که به اصطلاح «نرمال‌سازی متن[1]» نامیده می‌شود. این پیش‌پردازش‌ها در متون فارسی می‌تواند

1 Normalization

شامل، یک‌دست کردن فاصله‌ها، نشانه‌گذاری‌های درون متن، یکسان کردن کاراکترهای استفاده شده در متون، یکسان کردن روش اتصال وندهای گوناگون به ستاک، اصلاح غلط‌های املایی، ارتباط دادن کلمات چندواژه‌ای و یکسان در نظر گرفتن آن‌ها و....باشد؛ برای نرمال‌سازی متن، می‌توان از نرم‌افزارهای تعیین شده در این حوزه استفاده کرد. سپس، متنی که عملیات پیش‌پردازش روی آن صورت پذیرفته است، به اصطلاح نرمالایز شده، وارد بخشی می‌شود که برچسب‌گذاری اجزای واژگانی کلام[1] روی آن انجام می‌شود تا مقولهٔ واژگانی کلمات و نشانه‌های نوشتاری تشکیل‌دهندهٔ متن (شامل: اسم، صفت، فعل، قید، علائم سجاوندی و....) مشخص شود؛ این بخش را می‌توان با استفاده از نرم‌افزارهای مربوطه یا به صورت دستی انجام داد. البته می‌توان ابتدا از نرم‌افزارهای برچسب‌دهی به اجزای کلام استفاده کرد و سپس به صورت دستی، برچسب‌ها را چک کرد. پس از این مرحله، باید متن را مجدداً پالایش کرد؛ به این صورت که تمام واژگانی که بار معنایی واژگانی ندارند، و به اصطلاح واژگان دستوری خوانده می‌شوند (مانند حروف اضافه، حروف تعریف، حروف ربط) حذف گردند، همچنین از آنجائیکه اعداد و علائم سجاوندی (مانند نقطه، ویرگول، دو نقطه، پرانتز و...) کلمات کلیدی محسوب نمی‌شوند، این علائم و اعداد نیز باید حذف شوند. هولت (۲۰۰۳) معتقد است اکثر کلیدواژه‌ها از مقولات زیر هستند:

- صفت+ اسم (در فارسی این ترتیب به صورت«اسم+صفت» است)
- اسم+ اسم
- اسم (به تنهایی)

بنابراین، در مدل پیشنهادی نیز، قیدها حذف می‌شوند و آنچه باقی می‌ماند مجموعه‌ای است از اسم‌ها و صفت‌ها، افعال و اختصارات. در برچسب‌گذاری ماشینی اجزای کلام، می‌توان از نرم‌افزارهایی استفاده کرد که فراوانی کلمات را کنار برچسب آن‌ها نشان می‌دهد و بنابراین، در نهایت فهرستی از اسم‌ها و صفت‌ها، افعال و اختصارات همراه با فراوانی آن‌ها خواهیم داشت که می‌توان ۳۰ کلمه اول را بررسی کرد و از میان آن‌ها ۱۰ یا ۱۵ عبارت را که می‌تواند ترکیب دو اسم، یا اسم و صفت، اسم به تنهایی یا فعل و اختصارات باشد و فراوانی بالایی در متن دارند، به عنوان کلیدواژه انتخاب کرد. این مدل را می‌توان ابتدا روی متن نسبتاً کوتاهی آزمایش کرد و کلیدواژه‌ها را دستی کنترل کرد؛ چنانچه روش پیشنهادی از صحت بالایی برخوردار بود، می-

1 Part of Speech (POS) tagging

توان این روش را برای متون بلندتر نیز اجرا کرد.

۳- نتیجه‌گیری

در پژوهش حاضر، پس از تعریف «استخراج کلیدواژه» و کاربردها و اهداف این امر، به بررسی برخی از پژوهش‌های انجام شده در زمینهٔ استخراج کلیدواژه‌ها از متون فارسی پرداخته شد. در نهایت مدلی برای استخراج کلیدواژه معرفی شد که مبتنی بر دانش زبان‌شناسی رایانشی و آمار است. این مدل را می‌توان ابتدا روی داده‌های محدود و سپس متون گوناگون، به کار برد.

منابع

اسماعیلی آدرگانی، زهرا و احمد فراهی. (۱۳۹۲). استخراج کلیدواژه‌های مقالات فارسی بر اساس تحلیل معنایی، *هفتمین کنفرانس ملی مهندسی برق با محوریت انرژی‌های نو*. علی آباد کتول. دانشگاه آزاد اسلامی واحد علی آباد کتول. https://www.civilica.com/Paper-NEEREC07-NEEREC07_031.html

راد، فرهاد.، پروین، حمید.، دهباشی، آتوسا. و بهروز مینایی. (۱۳۹۵). ارائه روشی جدید برای شاخص گذاری خودکار و استخراج کلمات کلیدی برای بازیابی اطلاعات و خوشه بندی متون. *پردازش علائم و داده‌ها*. ۱۰۰-۸۷.

گزنی، علی. (۱۳۸۵). استخراج خودکار عبارت‌های کلیدی از متون مقاله‌های فارسی. *کتابداری و اطلاع‌رسانی*. ۹ (۳). ۱۰۸-۹۷.

معادی، معین و کاظم فولادی قلعه. (۱۳۹۴). ارائه روشی برای استخراج کلمات کلیدی در زبان فارسی. *دومین کنفرانس بین المللی و سومین همایش ملی کاربرد فناوری های نوین در علوم مهندسی*.

Beliga, Slobodan. (2014). Key word extraction: a review of methods and approaches. University of Rijeka. 1-9.

Hulth, Anette. (2003). Improved automatic key word extraction given more linguistic knowledge. *Proceedings of the conference on Empirical methods in natural language processing* . 216-223

Kaur, Jasmeen., and Gupta, Vishal. (2010). Effective approaches for

extraction of keywords. *IJCSI International Journal of Computer Science Issues*, 7 (6). 144-148.

Siddiqi, Sifatullah. (2015). Key word and key phrase extraction techniques: A literature review. International journal of computer applications. (0975-8887). 109 (2).

Turney. Peter D. (2000). Learning algorithms for key phrase extraction. *Information retrieval*. 2 (4). 303-336.

مرکز منطقه ای اطلاع رسانی علوم و فناوری

ISC
Islamic World Science Citation Center

پایگاه استنادی علوم جهان اسلام

مجموعه مقالات

دومین همایش ملی

پژوهش های کاربردی در زبان شناسی رایانشی
(با محوریت خط و زبان فارسی)

شیراز - مهر ۱۳۹۸

مجموعه مقالات دومین همایش ملی پژوهش های کاربردی در زبان شناسی رایانشی (با محوریت خط و زبان فارسی)

RICeST

ISC
Islamic World Science Citation Center

Proceeding of

The 2nd National Conference on Applied Research in Computational Linguistics
(with Emphasis on Persian Script and Language)

Shiraz - October 2019