دانشکده علوم تربیتی و روانشناسی

گروه علم اطلاعات و دانش‌شناسی

محاسبه بار اطلاعاتی واژه در متون علمی فارسی براساس شاخص آنتروپی نظریه اطلاعات

استاد راهنما

دکتر محمدرضا داورپناه

استاد مشاور

دکتر رحمت‌الله فتاحی

دانشجو

اعظم بیگ‌لو

مهرماه ۱۳۹۲

تعهد نامه

عنوان پایان نامه :

محاسبه بار اطلاعاتی واژه در متون علمی فارسی براساس شاخص آنتروپی نظریه اطلاعات

اینجانب **اعظم بیگ‌لو** دانشجوی دوره کارشناسی‌ارشد **رشته کتابداری و اطلاع رسانی** دانشکده علوم تربیتی و روانشناسی دانشگاه فردوسی مشهد تحت راهنمایی **دکتر محمدرضا داورپناه** متعهد می شوم:

- تحقیقات در این رساله توسط اینجانب انجام شده و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است.
- مطالب مندرج در این رساله تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی به جایی ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه فردوسی مشهد است و مقالات مستخرج با نام "دانشگاه فردوسی مشهد" و یا "Ferdowsi University of Mashhad" به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی رساله تاثیرگذار بوده‌اند در مقالات مستخرج از آن رعایت شده است.
- در کلیه مراحل انجام این رساله، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده، اصل رازداری، ضوابط و اصول اخلاق انسانی رعایت شده است.

تاریخ

امضای دانشجو

چکیده

هدف: هدف عمده این پژوهش، محاسبه میزان بار اطلاعاتی واژه‌های متون علمی فارسی و بررسی رابطه برخی ویژگی‌های واژه و بار اطلاعاتی آن بر مبنای مقیاس آنتروپی شانون است.

روش: پژوهش حاضر با روش تحلیل محتوا و در جامعه آماری شامل ۷۵۲ مقاله برگرفته از فهرست مجلات علمی پژوهشی در سال ۱۳۸۸ صورت پذیرفت. نمونه پژوهش شامل ۳۲۰ مقاله بود که با توجه به گستردگی آن در هر حوزه تنها ۱۰ درصد از مقالات به صورت تصادفی انتخاب و مورد بررسی قرار گرفت. بدین ترتیب، در این پژوهش در مجموع ۱۱۱۱۱۹ واژه بررسی شد که ۶۱۱۷۰ واژه مربوط به حوزه ادبیات و علوم انسانی، ۸۲۷۷ واژه مربوط به حوزه علوم پایه، ۱۴۱۶۲ واژه مربوط به حوزه فنی و مهندسی، و ۲۷۵۱۰ واژه مربوط به حوزه کشاورزی و منابع طبیعی است.

یافته‌ها: پژوهش حاضر نشان داد بار اطلاعاتی واژه با احتمال رخداد آن رابطه‌ای معکوس دارد که این احتمال رخداد با افزایش تعداد حالات ممکن واژه، افزایش یافته و بنابراین، واژه اطلاعات کمتری منتقل مـــــی‌نماید. علاوه بر میزان اطلاعات واژه، میزان اطلاعات متن نیز قابل محاسبه است که این اطلاعات در متون با افزایش آنتروپی کاهش می‌یابد. مقایسه آنتروپی در چهار حوزه ادبیات و علوم انسانی، علوم پایه، فنی و مهندسی، و کشاورزی و منابع طبیعی نشان داد، حوزه‌های مختلف علمی در میزان اطلاعاتی که انتقال می‌دهند یکسان نیستند و حوزه علوم انسانی بیشترین میزان آنتروپی و کمترین میزان اطلاعات را نسبت به سایر حوزه‌ها داراست.

کلیدواژه‌ها: نظریه اطلاعات، آنتروپی، بار اطلاعاتی واژه

فهرست مندرجات

فهرست جداول

فهرست شکل‌ها

فصل اول

کلیات پژوهش

۱-۱. مقدمه

هر نظام اطلاع‌رسانی به منظور بازنمون اطلاعات به نمایه‌ای مناسب نیاز دارد. جستجوی بهینه اطلاعات به طور عمده متکی بر نمایه‌سازی است و بدیهی به نظر می‌رسد که بدون وجود نمایه، بازیابی اطلاعات غیرممکن شده و با شکست مواجه می‌شود.

هدف از فرآیند نمایه‌سازی، جایابی اطلاعات به طور دقیق و باصرفه است. اختصاص واژه‌نماهای[1] خاص به مدارک، واژه‌نماهایی که بیشترین اطلاعات را نسبت به سایر واژگان در بافت متن انتقال می‌دهند، به همین منظور صورت می‌پذیرد. مؤسسه استاندارد ملی آمریکا فرایند نمایه‌سازی را اینگونه تعریف می‌کند: تحلیل محتوای اطلاعاتی رکوردهای دانش،که شامل انتخاب مفاهیم قابل نمایه شدن در مدرک و بیان این مفاهیم به زبان نظام نمایه‌سازی است (Borko & Bernier, 1978). بنابر این تعریف، انتخاب واژه‌نما در فرآیند نـمایه‌سازی کلیدی‌ترین موضوع است.

در سال‌های اخیر با ورود فرآیند نمایه‌سازی ماشینی به حوزه بازیابی اطلاعات روش‌های انتخاب کلیدواژه و بازنمون متون اهمیت خاصی یافته است. در واقع با افزایش حجم پیکره متون و اطلاعاتی که باید بازیابی شوند، روش‌های به کار گرفته شده برای نمایه‌سازی لازم است دقیق‌تر و مبتنی بر فنون ریاضی و آماری مورد اطمینانی باشد.

نظام‌های نمایه‌سازی خودکار برای استخراج لغات متن به طور معمول از روش‌های زیر استفاده می‌کنند:

۱. روش زبانشناختی که در آن، به کمک تجزیه تحلیل‌های ریخت‌شناسی و ساختار نحوی و معنایی متن مدارک، اقدام به استخراج واژه‌ها یا اصطلاحات نمایه‌ای می‌شود.
۲. روش آماری که در آن، معنی هر مفهوم واحد در مدرک با حضور آن در جای جای مدرک ارتباط دارد. لذا واژه‌های متن مدارک شمرده می‌شوند و ارتباطشان مورد ارزش‌گذاری قرار می‌گیرد تا براساس آن‌ها واژه‌های دارای بار معنایی به عنوان اصطلاحات نمایه‌ای انتخاب شوند.

1. index term

۳. روش مبتنی بر احتمالات که در آن، نظریه احتمالات برای مدلسازی ریاضی مراحل بازیابی مورد استفاده قرار می‌گیرد (آقابخشی، ۱۳۸۶).

در هر سه روش فوق آن‌چه با اهمیت است، نوعی الگوریتم وزن‌دهی به کلمات است. همان‌طور که گفته شد به منظور وزن‌دهی کلمات و ساختن سیاهه‌های مجاز و غیرمجاز واژگان، استفاده از مدل‌ها و فرمول‌های آماری نیز کاربرد دارد. یکی از روش‌های مدلسازی آماری، استفاده از کمیت آنتروپی نظریه ریاضی اطلاعات[1] شانون[2] می‌باشد. این نظریه شاخه‌ای از نظریه آماری علوم ارتباطی است و شیوه کمّی جدیدی برای اندازه‌گیری محتوای اطلاعاتی پیام‌ها و ابداع کارآمدترین رمزها برای انتقال آن‌ها به دست می‌دهد. هدف اصلی شانون دست یافتن به شیوه‌ای بود که کارایی کانال ارتباطی را از نظر انتقال درست و کامل اطلاعات به حداکثر برساند. این نظریه به طور عمده ناظر به مسئله تعیین حداکثر ظرفیت یک کانال یا یک مجرا برای انتقال پیام‌هاست (حری،۱۳۸۱). نظریه ریاضی اطلاعات در شاخه‌های مختلف مهندسی الکترونیک، فیزیک و ترمودینامیک به کار رفته است، اما از آن جا که می‌توان بین مهم‌ترین شاخص آن یعنی آنتروپی با اطلاعات ارتباط برقرار نمود، به منظور اندازه‌گیری اطلاعات در هر واحد معنایی پیام می‌تواند مورد استفاده قرار گیرد. در واقع علاوه بر کار، حرارت و انرژی، مفهوم اطلاعات با مفهوم قدیمی آنتروپی پیوند خورده است: اطلاعات عبارت از نظم یا نگانتروپی است. در اینجا، آنتروپی به فقدان اطلاعات مشاهده‌گر درباره نظامی که مورد بررسی قرار می‌دهد تبدیل می‌شود؛ حداکثر آنتروپی، حداکثر نادانی است. به عبارت دیگر، آنتروپی در برداشت بسیار رایج خود نه تنها بی‌نظمی یا نبود سازمان در یک نظام فیزیکی، بلکه کاهش اطلاعات مشاهده‌گر درباره موضوع مورد مشاهده خود را نیز می‌سنجد (نشاط، ۱۳۸۵).

می‌توان رابطه اطلاعات و آنتروپی را از منظر دیگری نگریست. سورین و تانکارد[3] (۱۳۸۴) معتقدند در نظریه اطلاعات، اطلاعات با آنتروپی در علوم فیزیکی، که مقیاس، درجه اتفاقی بودن است خیلی شباهت دارد؛ آنتروپی، عدم اطمینان یا از هم گسیختگی یک وضعیت است و در یک پیام خیلی نظم‌یافته درجه اتفاقی بودن، عدم اطمینان یا انتخاب، بالا نیست. در این حالت، اطلاعات اندک است، زیرا هر بخش از پیام که هنگام دریافت

1. Mathematical Information Theory
2. CludeShanon
4. Severin&Tankard

از بین برود، احتمال زیادی وجود دارد که گیرنده بتواند حدس بزند چه چیزی حذف شده یا از دست رفته است. بنابراین می‌توان آنتروپی را وضعیتی دانست که در صورت نبودن اطلاعات رخ می‌نماید. هرچه در یک سیستم عدم اطمینان بیشتر باشد، پیش‌بینی‌پذیری کمتر خواهد بود و اطلاعاتی که منتقل می‌شود بالاتر است، در این صورت با وجود اطلاعات، آنتروپی اندکی در سیستم مشاهده خواهد شد.

طبق آن‌چه گفته شد شاخص آنتروپی نیز می‌تواند به عنوان روشی در جهت وزن‌دهی به واژگان در فرآیند نمایه‌سازی به کار رود. استفاده از شاخص آنتروپی جایگزین روش‌های نمایه‌سازی واژگان براساس آمارهای فراوانی واژه شده است. روش‌های مبتنی بر فراوانی سطحی هستند و مفهوم اصلی متن را منعکس نمی‌کنند، در حالیکه مدل‌هایی مبتنی بر آنتروپی دقت بالایی را گزارش کردند (Fragos, Maistros & Skourlas, 2005; Lv & Liu, 2005).

۱-۲. بیان مسأله

پیشرفت روزافزون نظام‌های ذخیره و بازیابی اطلاعات و نیاز به بهینه نمودن بازیابی اطلاعات از مدارک متنی باعث شده است تا در سال‌های اخیر توجه گسترده‌تری معطوف به فنون و رویکردها در زمینه نظام‌های نمایه‌سازی خودکار شود. پیامد استفاده از نظام‌های نمایه‌سازی مبتنی بر زبان طبیعی انعکاس انواع کلمات در نمایه است. از آن‌جا که تمام واژگان در متن ارزش و بار اطلاعاتی یکسانی ندارند استفاده از روش‌هایی که کلمات مهم را از کلمات بی‌اهمیت تشخیص دهد همیشه در این حوزه مورد توجه بوده است. گروهی از واژگان زبان طبیعی (مانند حروف تعریف، حروف ربط، حروف اضافه، و برخی از افعال) سهم معنایی یا دستور زبانی بسیار پایینی دارند. به عبارت دیگر، حشو ویژگی بارز متون زبان طبیعی است که به منظور جلوگیری از اختلال در درک پیام متن به کار می رود. از سوی دیگر، بسامد واژه به تنهایی، به منظور اختصاص کلیدواژه‌های موضوعی مدارک چندان قابل اعتماد به نظر نمی‌رسد (داورپناه و بلندیان، ۱۳۸۶). علاوه بر این، اگر فرایند استخراج کلیدواژه بدون توجه به بار اطلاعاتی و وزن معنایی کلمه انجام پذیرد، علاوه بر حجیم شدن پایگاه واژگان نمایه، ریزش کاذب و بازیابی منابع نامرتبط نیز دور از انتظار نخواهد بود، چرا که واژگان بار اطلاعاتی یکسانی ندارند. در واقع هر کلمه یا ترکیب به یک میزان اطلاع‌دهنده[1] نیست، یعنی بسیاری از ترکیبات و

1. informative

کلمات نباید در نظام‌های نمایه سازی به عنوان واژه‌نما انتخاب شوند. نظریه اطلاعات ناظر بر روابط میان این سه عامل است: چگونگی رمزگذاری پیام‌ها، وجود اختلال (یعنی هرگونه شرایطی که بخشی از علائم را دگرگون کند)، و ظرفیت کانال. به این ترتیب، و با توجه به عوامل یاد شده، نظریه اطلاعات با ارائه شاخصی به نام آنتروپی به اندازه‌گیری اطلاعات یک متغیر تصادفی می‌پردازد. این متغیر تصادفی می‌تواند واحدی از یک متن (حرف، کلمه، جمله، و مانند آن) باشد. با استفاده از آنتروپی می‌توان میزان بار اطلاعاتی[1] یک واژه را به عنوان متغیری تصادفی اندازه گرفت و کلمات با آنتروپی بالا را نادیده گرفت. به همین ترتیب می‌توان سیاهه‌ای از واژگان غیرمجاز ساخت. مطالعات نیز نشان داده است که استفاده از شاخص آنتروپی نسبت به سایر معیارهای نحوی برای تشخیص و شناسایی واژگان کارکردی و ساخت سیاهه واژگان غیرمجاز مفیدتر است (Melamed, 1997). بنابراین می‌توان گفت شاخص آنتروپی طرح وزن‌دهی مناسبی برای شناسایی واژگان کم‌بار و پربار اطلاعاتی است. در واقع آن‌چه در یک روش وزن‌دهی به واژه قابل توجه است تعیین میزان اهمیت واژگان می‌باشد که بر اساس آن می‌توان واژگان مهم را از واژگانی که در متن دارای اهمیت کمتری هستند متمایز نمود. واژگان کم‌اهمیت در یک طرح وزن‌دهی به طور معمول به عنوان واژه‌نما انتخاب نمی‌شوند، این واژگان همان واژگان غیرمجاز هستند که در نظریه اطلاعات و با استفاده از آنتروپی نیز قابل شناسایی هستند. واژگان غیرمجاز کمترین میزان اطلاعات و بنابراین بیشترین میزان آنتروپی را دارا هستند؛ واژه‌های مهم یک متن نیز به همین ترتیب تعیین می‌شوند. با توجه به آنچه گفته شد، مسأله اساسی این پژوهش آن است که بر مبنای آنتروپی نظریه اطلاعات بار اطلاعاتی کلمات در متون علمی فارسی چگونه است؟

۱-۳. اهمیت و ضرورت پژوهش

با گسترش منابع اطلاعاتی، بیش از پیش لزوم توجه به فنون نمایه‌سازی و به کارگیری شاخص‌های دقیق‌تر و کمّی‌تری در این زمینه احساس شده است. در زبان فارسی نیز با عنایت به لزوم بهبود نظام‌های بازیابی اطلاعات و نمایه‌سازی مدارک و نیز برخی ویژگی‌های خاص خط و زبان فارسی بررسی مسائل مختلف در این حوزه شایان توجه است؛ حال آن‌که در زبان فارسی کمتر به موضوع وزن‌دهی واژگان پرداخته شده است. لزوم پژوهشی در راستای تعیین میزان اهمیت واژه، می‌تواند آغازی بر انجام پژوهش‌ها در حوزه وزن‌دهی به واژه باشد.

1. information content

۱-٤. هدف‌های پژوهش

هدف اصلی این پژوهش سنجش میزان آنتروپی واژه در متون علمی و تخصصی زبان فارسی است. سایر هدف‌های پژوهش حاضر را می‌توان به این شرح برشمرد:

۱. محاسبه بار اطلاعاتی واژگان متن و شناسایی واژگان کم بار اطلاعاتی (واژگان غیرمجاز)
۲. مقایسه کلمات کلیدی اختصاص یافته به مقاله با کلمات دارای بار اطلاعاتی بالای استخراج شده از متن
۳. بررسی رابطه بار اطلاعاتی یک واژه با احتمال رخداد آن در متن
۴. تعیین رابطه بار اطلاعاتی واژه با تعداد حالات ممکن آن (شکل‌های مختلف واژه شامل هم‌خانواده‌ها و هم‌ریشه‌ها)
۵. بررسی رابطه بار اطلاعاتی واژه با طول واژه
۶. مقایسه میزان اطلاعات متون حوزه‌های مختلف علمی
۷. بررسی اطلاعات متن و رابطه آن با آنتروپی

۱-۵. فرضیه‌های پژوهش

۱. میزان بار اطلاعاتی واژه متن با احتمال رخداد آن رابطه معکوس دارد.
۲. هرچه تعداد حالات ممکن یک واژه (مترادفات) کمتر باشد احتمال وقوع آن واژه نیز کمتر است و حاوی اطلاعات بیشتری خواهد بود.
۳. هرچه میزان آنتروپی متن بیشتر باشد میزان حضور اطلاعات در متن کمتر است.
۴. مقدار اطلاعات متون در حوزه‌های مختلف علمی متفاوت است.
۵. بین طول کلمه و بار اطلاعاتی آن رابطه وجود دارد.
۶. واژه‌های عمومی در همه متون با بسامدی مشابه رخ می‌دهند.
۷. هرچه یک واژه در همه متون بیشتر تکرار شود، اهمیت کمتری دارد.

۸. کلمات کلیدی اختصاص‌یافته به مقاله با کلمات پربار اطلاعاتی استخراج شده از متن منطبق است[۱].

۱-٦. تعاریف مفهومی و عملیاتی

در این بخش به معرفی برخی از مهم‌ترین مفاهیم پژوهش حاضر پرداخته می‌شود.

۱-٦-۱. آنتروپی (Entropy): آنتروپی شاخصی است که برمبنای آن اطلاعات کمیت‌پذیر می‌شود. رابطه اطلاعات با آنتروپی را به این صورت می‌توان شرح داد که اطلاعات، مقیاس عدم اطمینان یا آنتروپی در یک موقعیت است، هرچه عدم اطمینان (آنتروپی) بیشتر باشد، اطلاعات کمتر خواهد بود. به عبارت دیگر، وقتی موقعیّتی کاملا قابل پیش‌بینی است، هیچ اطلاعاتی وجود ندارد. این وضعیت را استحکام (نگانتروپی) می‌گویند (لیتل‌جان[۲]، ۱۳۸۴). شانون تعریف کرد اطلاعات I در پیام x با احتمال آن تناسب دارد.

$$I(x)= -Log_2P(x)\ P(x)$$

این فرمول مبتنی بر تعدادی از بیت‌هاست که برای انتقال پیام x در یک سیستم دودویی استفاده می‌شود. در واقع محتوای اطلاعات شانون $I(x)$ سنجش محتوای اطلاعاتی رخداد x_i است. برای مثال آنتروپی ۲۷ حرف زبان انگلیسی هنگامی که یک حرف به طور تصادفی از مدارک استخراج می‌شود محاسبه شده است. هنگامی که متغیر تصادفی x = z باشد آنتروپی برابر با ۱۰/۴ بیت و در صورتی که متغیر x = e برابر با ۳/۵ بیت است.

به این ترتیب آنتروپی در این پژوهش، مقیاسی برای اندازه‌گیری اطلاعات یک پیام (واژه یا متن) است و رابطه آن با اطلاعات رابطه‌ای معکوس می‌باشد که حاصل‌ضرب احتمال رخداد یک واژه در لگاریتم احتمال رخداد آن واژه در متن است و با توجه به علامت منفی در معادله شانون، مقدار اطلاعات عددی مثبت به دست می‌آید.

۱-٦-۲. احتمال (Probability): مفهوم احتمال برگرفته از نظریه احتمالات است. نقطه شروع نظریه احتمال انجام آزمایش‌هایی است که منجر به پیشامدهایی می‌شود. در این صورت هر رخداد را می‌توان به عنوان یک پیشامد در نظر گرفت. فرض بر این است که قادر به تشخیص پیشامدهای ممکنی که می‌تواند رخ دهد

---

۱. فرضیه‌های مطرح شده مربوط به متون علمی فارسی است.

8. Little John

هستیم، مجموعه همه پیشامدهای ممکن را فضای نمونه می‌نامند. با توجه به فضای احتمال X هر پیشامدی احتمال رخداد معینی دارد. احتمال مربوط به x_i را با $P(x_i)$ نشان داده و آن را توزیع احتمال می‌نامند (لوبه[9]، ۱۳۸۰). در واقع، احتمال یک پیشامد عددی است بین صفر و یک و شامل صفر و یک. اگر امکان وقوع یک پیشامد غیرممکن باشد احتمال آن پیشامد صفر است و اگر امکان وقوع پیشامدی حتمی باشد در این صورت احتمال آن پیشامد یک است.

۱-٦-۳. واژه (Word): واژه واحد آوایی مرکبی است که از ترکیب یک یا چند تکواژ حاصل می‌شود و به تنهایی در معنی و مفهومی مستقل به کار می‌رود. اندازه و تعداد تکواژهایی که یک واژه را می‌سازند همیشه یکسان نیست؛ در زبان فارسی یک واژه ممکن است مرکب از یک تا شش تکواژ باشد (باقری، ۱۳۶۷). درواقع واژه یک نماد زبانی است که دارای کارکرد خاصی در فرآیند برقراری ارتباط بوده و معنی و مفهوم آن وابسته به متن است. در این تحقیق، کلیه کلمات دارای بار معنایی و بدون بار معنایی به عنوان واژه در نظر گرفته می‌شود. روش تشخیص و جداسازی واژه‌ها نیز معیارهای مطرح در پژوهش‌های داورپناه و بلندیان (۱۳۸۶)، سنجی و داورپناه (۱۳۸۸) و برگرفته از قواعد مندرج در کتاب وحیدیان کامیار (۱۳۷۹) است.

۱-٦-٤. بار اطلاعاتی (Information content): در نظریه شانون، اطلاعات هر واحد معنایی قابل اندازه‌گیری است، زیرا این اطلاعات چیزی جز مجموعه‌ای از نمادها و رمزها و احتمال وقوع رویدادها نیست. در این الگوی اندازه‌گیری، تعداد نمادها شاخص اندازه‌گیری درجه احتمال رویداد تلقی می‌شود (حری، ۱۳۸۷). در این تحقیق منظور از بار اطلاعاتی، میزان آنتروپی آن واژه است که از طریق معادلات ذکر شده قابل اندازه‌گیری است. البته آنتروپی به معنای عدم اطلاعات مورد توجه قرار می‌گیرد.

۱-٦-۵. واژگان کلیدی (key words): کلیدواژگان، واژگانی هستند که نشان‌دهنده هدف و منظور اصلی نویسنده یک متن می‌باشند. در این پژوهش منظور از واژگان کلیدی، واژگان ارائه شده توسط نویسنده تحت عنوان کلیدواژه است.

9. Lubbe

۱-٦-٦. طول واژه (word length): تعداد حروف یا نویسه‌های یک واژه به عنوان طول واژه تعریف شده است (Jacroux, 2004). ما نیز در این پژوهش طول هر واژه را معادل تعداد حروف آن در نظر گرفتیم.

۱-٦-۷. حالات ممکن (possible status): حالات ممکن یک واژه، شکل‌هایی از یک واژه است که با یک ریشه مشترک در متن وقوع می‌یابد. در پژوهش حاضر همه هم‌خانواده‌های یک واژه و کلیه شکل‌های دستور زبانی آن مانند اسم، صفت، فعل، قید و ... در شمار حالات ممکن واژه قرار گرفتند.

فصل دوم

مبانی نظری و پیشینه پژوهش