

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِیْمِ



عنوان درس :

استاد مربوطه :

دانشجو :



عنوان مقاله

بررسی روش های پردازش زبان طبیعی در خلاصه سازی متن

امروزه افزایش بیش از پیش حجم اطلاعات موجود در وب باعث پیچیده شدن دسترسی به داده ها شده است. یکی از راه حل های کاهش این پیچیدگی استفاده از الگوریتم های خلاصه سازی و طبقه بندی اطلاعات می باشد. بطور کلی خلاصه سازی متن عبارت است از خلاصه سازی متن اصلی به شکل کوتاهتر با حفظ محتوای اطلاعاتی و مفهوم کلی آن. از آنجا که انجام این امر بصورت دستی امری غیر ممکن می باشد، لزوم ایجاد الگوریتم های دسته بندی خودکار بسیار نمایان می باشد. در این تحقیق به روش های خلاصه سازی متن که می توانند به خلاصه سازی استخراجی و چکیده طبقه بندی شوند اشاره شده است. همچنین در این مقاله به بررسی چالش های پیشرو در خلاصه سازی خودکار شامل انتخاب جملات مهم ، پاراگراف های مهم و غیره از اسناد اصلی و الحاق آنها در شکل کوتاهتر پرداخته می شود. اهمیت جملات مبتنی بر شکل های آماری و زبان شناسی جملات تصمیم گیری می شود.

یک روش خلاصه سازی چکیده شامل درکی از متن اصلی و ذکر چکیده آن در چند کلمه است که از روش های زبان شناسی برای آزمایش و تفسیر متن و سپس یافتن مفاهیم جدید و اصطلاحات جدید برای توصیف بهتر آن با ایجاد متن جدید کوتاهتر استفاده می کند که مهم ترین اطلاعات را از سند متن اصلی نقل می کند.

بزرگترین ایراد برای خلاصه سازی متن خلاصه سازی محتوا از یک سری منابع متنی و نیمه سازماندهی شده است که شامل مبنای داده ها و متون وب در روش صحیح است. نرم افزار خلاصه سازی متن باید خلاصه موثر را در زمان کمتر و با کمترین فراوانی فراهم کند.



مقدمه

یکی از شاخه های مرتبط به خلاصه سازی متون، متن کاوی است که به معنای کشف اطلاعات جدید به وسیله استخراج اطلاعات به صورت اتوماتیک و بر اساس منابع مختلف می باشد. در متن کاوی تعداد زیادی متن تحلیل شده و الگوهای با ارزشی که در متن مخفی است یافته می شود. بخش متن کاوی در سالهای اخیر شامل کشف ارتباط بین کلمات، طبقه بندی اسناد و خلاصه سازی شده است که در این مقاله به مقوله خلاصه سازی پرداخته می شود، خلاصه سازی یکی از کاربردهای پردازش متن است.

فرآیند جمع آوری و کوتاه کردن یک یا چند سند توسط ماشین، به طوری که مفاهیم و نکات اصلی آنها حفظ شود، خلاصه سازی نام دارد. خلاصه سازی را می توان یک فرآیند تصمیم گیری در نظر گرفت زیرا ارزش جملات بر اساس یک سری معیار سنجیده می شود و باید تعدادی از جملات انتخاب شوند. متن خلاصه شده باید مختصر، روان، پیوسته، آموزنده، خوانا و حاوی نکات اصلی اسناد و همچنین فاقد ایرادهای دستور زبانی، حشو و... باشد. خلاصه سازی خودکار متون بخشی از دانش یادگیری ماشین و پردازش زبان طبیعی است و ایده اصلی آن یافتن زیر مجموعه ایی از داده هایی است که اطلاعات کلی را در بر دارند. این علم در پردازش زبان طبیعی بسیار محبوب می باشد، زیرا از قابلیت های آن می توان در سیستم های پرسش و پاسخ، موتورهای توصیه شخصی، خلاصه سازی محتوای صفحات وب استفاده کرد. هدف تمامی سیستم های خلاصه ساز خودکار گردآوری خلاصه ایی از اطلاعات مطابق با نیاز کاربر می باشد، به طوری که خلاصه تولیدی از نظر کیفیت به خلاصه های انسانی شبیه باشد.

نرخ تولید اطلاعات بر روی شبکه گسترده وب به صورت روز افزون در حال افزایش می باشد بنا به گفته مرکز صنعت اطلاعات جهانی، هر ساله ۳۰ درصد به حجم اطلاعات وب اضافه می گردد. افزایش روز افزون منابع اطلاعاتی در بستر اینترنت، جستجو و یافتن اطلاعات مورد نیاز را مشکل کرده است. از طرفی مشکل محققان در عصر حاضر دسترسی دقیق به اطلاعات مورد نیازشان در کوتاه ترین زمان ممکن از میان حجم انبوهی از اطلاعات سطح وب می باشد.