نوزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران
دانشکده مهندسی کامپیوتر، دانشگاه شهید بهشتی، تهران، ۱۵-۱۳ اسفندماه ۱۳۹۲

ارائه رویکرد بدون ناظر در محاسبه شباهت معنایی اسناد متنی کوتاه

مجید محبی[۱]، علیرضا طالب‌پور[۲]

[۱] کارشناس ارشد نرم افزار، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی، تهران،
ma.mohebbi@mail.sbu.ac.ir

[۲] استادیار، دانشکده مهندسی برق و کامپیوتر، دانشگاه شهید بهشتی، تهران،
talebpour@sbu.ac.ir

چکیده

بخش قابل توجهی از اطلاعات در دسترس، در پایگاه داده‌های متنی ذخیره شده است. به طور معمول تنها بخش کوچکی از اسناد در دسترس، برای یک فرد یا کاربر مناسب است. از اینرو تولید پرس‌وجوی مناسب سندی، برای تحلیل و استخراج اطلاعات مفید از اسناد متنی، مشکل است. این امر اهمیت موضوع شباهت اسناد متنی را دو چندان می‌کند.

انواع مختلفی از روشهای تطبیق لغوی، برای تعیین شباهت بین اسناد ارائه شد که تا یک حد خاصی موفق عمل می‌کردند ولی قادر به تشخیص شباهت معنایی بین دو متن نبودند. از اینرو، رویکردهای شباهت معنایی مطرح شد که از میان آنها می‌توان روشهای مبتنی بر پیکره و روشهای مبتنی بر پایگاه دانش مانند وردنت را نام برد. هدف ما این است که در حوزه‌ی مدل‌های شباهت معنایی و مبتنی بر پایگاه دانش وردنت، با ارائه یک رویکرد بدون ناظر، میزان شباهت بین اسناد انگلیسی را با دقت مناسبی محاسبه کنیم؛ برای این منظور، از مدل گرافی بهره می‌بریم و برای ارزیابی، از مجموعه داده‌ی Microsoft Research Paraphrase Corpus **استفاده می‌کنیم. ارزیابی انجام شده، عملکرد مناسب رویکرد پیشنهادی را نشان می‌دهد.**

کلمات کلیدی

شباهت معنایی، شباهت اسناد، معیارهای شباهت، وردنت، تئوری گراف.

۱- مقدمه

رشد فزاینده اطلاعات یکی از ویژگی‌های دنیای امروز است. در اغلب موارد، اطلاعات مهم به فرم متن ذخیره شده است. برخلاف داده‌های عددی، متن اغلب بی‌نظم و برای بررسی کردن مشکل است. در این حجم عظیم داده‌های متنی، نیاز است تا اطلاعات مناسب و دقیق بازیابی شود و مرتبط‌ترین متون از میان مجموعه مستندات به کاربر ارائه گردد.

محاسبه‌ی شباهت اسناد یک کار مشترک در انواع مسائلی همچون خوشه‌بندی اسناد، تحلیل مرجع مشترک[۱]، جستجوی سند مشابه، توصیه‌های[۲] سند، تشخیص سرقت ادبی[۳]، فیلتر کردن سند و ... است.

یک سند متنی، دنباله‌ای از کلمات و علائم نقطه گذاری است که از قواعد دستوری زبان پیروی می‌کند و می‌تواند به هر طولی باشد. بسته به نوع تحلیلی که انجام می‌شود و با توجه به اهداف محقق، یک سند می‌تواند مقاله، کتاب، صفحه‌ی وب، ایمیل و غیره باشد. یک سند ممکن است حتی فقط شامل یک فصل، یک پاراگراف و یا حتی یک جمله باشد.

در بیشتر مدل‌های پیشین، یک جمله به سادگی به عنوان یک مجموعه‌ای از کلمات یا یک «کیسه‌ای از کلمات[۴]» در نظر گرفته می‌شد و ترتیب کلمات می‌توانست بدون تأثیر در تحلیل خروجی تغییر کند؛ به عبارت دیگر به ترتیب ظاهر شدن کلمات توجهی نشده و ساختار نحوی یک جمله یا پاراگراف، به عمد نادیده گرفته می‌شد.

سیستم‌های بازیابی پیشین، بدون اینکه درک درستی نسبت به مفاهیم موجود در پرس و جوی سندی و مجموعه اسناد داشته باشند، سعی می‌کردند تا الگوهای مشابه با پرس و جوی سندی را در میان انبوه اسناد موجود در مخزن بیابند. با ارائه روش‌های معنایی سعی شد تا محدودیت‌های روش‌های شباهت لغوی کنار گذاشته شود و با محاسبه رابطه معنایی بین لغات و متون، مرتبط‌ترین مطالب از نظر معنا بازیابی شود و از استخراج مطالب غیر مرتبط و نامناسب جلوگیری شود.