

روشی جدید برای خوشه‌بندی مستندات متنی براساس آنتولوژی

احسان عسگریان^۱، جعفر حبیبی^۱، شهرزاد معاون^۱ و حسین معین زاده^۲

^۱دانشگاه صنعتی شریف

^۲دانشگاه علم و صنعت

asgarian@ce.sharif.edu - jhabibi@sharif.edu - moaven@ce.sharif.edu - ho_moeenzadeh@comp.iust.ac.ir

چکیده - امروزه خوشه‌بندی نقش حیاتی در روشهای بازبایی اطلاعات برای سازمانبندی مجموعه های بزرگ مستندات متنی مانند وب، درون تعداد کمی خوشه معنادار دارد. معمولاً در خوشه‌بندی مستندات متنی با خوشه‌بندی ابعاد بسیار بالای فضای داده مواجه هستیم که انجام خوشه‌بندی به این شکل، مشکل بنظر می‌رسد. علاوه بر این، معمولاً با استفاده از الگوریتم‌های خوشه‌بندی متفاوت، نتایج متفاوتی بدست می‌آید و مستندات درون خوشه‌های بدست آمده نیز ارتباط معنایی لازم را با هم ندارند. برای حل این مشکلات ما هسته آنتولوژی مانند وردنت را به عنوان دانش پس زمینه، درون فرایند خوشه‌بندی مستندات متنی جایگذاری نمودیم. سپس به صورت تجربی روش جدید را با تکنیکهای خوشه‌بندی ساده برای مجموعه داده خبرگزاری رویتر مقایسه نمودیم. نتایج بدست آمده از خوشه‌بندی متن همراه با استفاده از دانش پس‌زمینه در مقایسه با خوشه‌بندی متن بدون دانش پس‌زمینه، بهبود داشته است.

کلید واژه- آنتولوژی، خوشه‌بندی، دانش پس‌زمینه، مفهوم.

۱- مقدمه

خوشه‌بندی می‌تواند به صورت بدون نظارت^۳ و بدون نیاز به مجموعه آموزشی اولیه خوشه‌بندی را انجام دهند.

برای انجام خوشه‌بندی لازم است تا مستندات متنی در قالب بردار داده نگهداری شوند تا بتوان به راحتی با استفاده از معیارهای مختلف اندازگیری فاصله بردارها، مستندات مشابه را پیدا نمود. برای نمایش برداری این نوع داده ها از مدل فضای برداری^۴ [۱۴] استفاده می‌شود. در این مدل سعی می‌شود تا یک مجموعه لغت مرجع از لغات درون مستندات نگهداری شود. سپس برای هر متن، برداری با تعداد مولفه برابر با تعداد کلمات این مجموعه لغت در نظر می‌گیریم. هر مولفه می‌تواند تعداد تکرار کلمه متناظر با کلمه درون مجموعه لغت مرجع را در متن نشان دهد یا اینکه هر مولفه به شکل دودویی وجود یا عدم وجود کلمه متناظر در مجموعه لغت مرجع، درون متن موجود را مشخص نماید. ولی با استفاده از این مدل، بردار فضای حالت مستندات

امروزه حجم زیادی از دانش بشر در غالب متن های الکترونیکی وجود دارد. پس یافتن اطلاعات مورد نظر از میان این کوه اطلاعات بسیار مشکل است. برای پاسخ گویی به این چالش روشهای دسته بندی^۱ و خوشه‌بندی^۲ متون پیشنهاد شده اند. هدف آنها سازمانبندی مجموعه بزرگ مستندات متنی مانند وب درون تعداد دسته ها یا خوشه های معنی دار نسبتاً کمی می باشد. از جمله کاربرد این روشها در بازبایی اطلاعات از پایگاههای داده، پروسه پیشنهاد راه حل‌های هوشمند تجارت، مدیریت سیستم های اطلاعاتی یا پورتال سازمانی می باشد [۱]. با رشد سریع علم و افزوده شدن صفحات وب و پیدایش ویکی ها و وبلاگها، ساخت مجموعه آموزشی جامع برای دسته بندی صفحات وب، بسیار مشکل است. پس استفاده از روشهای خوشه‌بندی داده می تواند رویکرد مناسبی باشد [۲]. روشهای