نخستین کنفرانس بین المللی پردازش خط و زبان فارسی
۱۵ و ۱۶ شهریور ۱۳۹۱
دانشگاه سمنان - دانشکده مهندسی برق و کامپیوتر

خوشه بندی جملات فارسی مبتنی برالگوریتم های هوش جمعی

مهدی بازقندی[۱]، قمرناز تدین تبریزی[۲] و مجید وفایی جهان[۳]
Mehdi_Bazghandi@ymail.com[۱]
Tadayon@mshdiau.ac.ir [۲]
VafaeiJahan@mshdiau.ac.ir[۳]

چکیده

خوشه بندی یکی از مسائل مهمی است که امروزه بسیاری از محققین در زمینه های مختلف به ان پرداخته اند. تا کنون الگوریتم های کلاسیک زیادی در این زمینه ارائه شده است. که اغلب این روش ها دارای ناپایداری بوده و همچنین پارامترهای ان ها محدود به انتخاب کاربر می باشد.از کاربردهای خوشه بندی می توان به خوشه بندی متون و اسناد در موضوعات خلاصه سازی متون و بازیابی اطلاعات یاد کرد. در خوشه بندی جملات یک متن برای مشخص شدن جملات مشابه , نمی توان از روش مشابه ان(دسته بندی متون مشابه)استفاده کرد . .بردارهایی به طول m و با مقادیر صفر بسیار زیاد پدید خواهد امد.برای حل این مشکل, روشی جدید مبتنی بر PSO برای خوشه بندی جملات یک متن معرفی شده است.به طوریکه به جای استفاده از فاصله اقلیدسی و فاصله کسینوسی ,از یک معیار جدید در محاسبه فاصله دو جمله استفاده شده است.معیاری که در ان, ارتباط معنایی کلمات با استفاده ازارتباطات انها در متن در نظر گرفته می شود. همچنین تعیین تعداد خوشه های بهینه یکی دیگر از کارهای انجام شده در این مقاله است.برای ارزیابی یک مجموعه از خبرهای ورزشی فارسی انتخاب شده است. نتایج حاصل از ارزیابی روش پیشنهادي نشان می دهند که استفاده از خوشه بندي PSO معنایی ,با تعیین تعداد خوشه های مطلوب, دقت بهتری را درخوشه بندی جملات در مقایسه با روش های دیگر.دارد.

کلمات کلیدی

خوشه بندی , الگوریتم PSO , بردارهای Contecxt- Vector ,شباهت معنایی

۱- مقدمه

پردازش داده، یکی از شاخص‌های بسیار مهم در دنیای اطلاعات است. خوشه بندی یکی از بهترین روش‌هایی است که برای کار با داده‌ها ارائه شده است. خوشه بندی قابلیت ورود به فضای داده و تشخیص ساختارش را امکان پذیر می‌نماید. لذا به عنوان یکی ازایده آل ترین مکانیزم‌ها، برای کار با دنیای عظیم داده‌ها محسوب می‌شود.

خوشه بندی، یافتن ساختاری در مجوعه‌ای از داده‌ها است که طبقه بندی نشده‌اند. به بیان دیگر می توان گفت که خوشه بندی قراردادن داده‌ها در گروه‌هایی است که اعضای هر گروه از زاویه خاصی شبیه یکدیگرند. در نتیجه شباهت بین داده‌های درون هر خوشه حداکثر و شباهت بین داده‌های درون خوشه های متفاوت حداقل می‌باشد. معیار شباهت در اینجا، فاصله بوده یعنی نمونه‌هایی که به یکدیگر نزدیکترند در یک خوشه قرار می‌گیرند. به عنوان نمونه در خوشه بندی اسناد دوری و یا نزدیکی داده‌ها متناسب با تعداد کلمه‌های مشترکی که در دو سند وجود دارد و یا در خوشه بندی سبد خرید مشتریان، فاصله بر اساس شباهت خرید تعیین می‌شود. لذا محاسبه فاصله بین دو داده در خوشه بندی بسیار مهم می‌باشد؛ زیرا کیفیت نتایج نهایی را دستخوش تغیر قرار خواهد داد. فاصله که همان معرف عدم تجانس است حرکت در فضای داده‌ها را می‌سازد وسبب ایجاد خوشه‌ها می‌گردد. با محاسبه فاصله بین دو داده می‌توان فهمید که چقدر این دو داده به هم نزدیک هستند و براین اساس در یک خوشه قرار داده می‌شود. توابع ریاضی مختلفی . برای محاسبه فاصله وجود دارند؛ فاصله افلیدسی، فاصله همینگ و...با افزایش حجم منابع متنی موجود در وب چالشی که و جود داشته است ان است که چگونه کاربران می توانند به اطلاعات مورد نیاز خود در اسرع وقت و با دقت بالا دسترسی داشته باشند.خوشه بندی می تواند راه حلی برای حل این مسئله باشد. .خوشه بندی می تواند نقش مهمی در انتخاب عنصرهای شایسته داشته باشد.با خوشه بندی اطلاعات می توان عنصرهای مشابه را دریک خوشه قرار داد و عنصرمحوری را به عنوان نماینده خوشه انتخاب کرد.خوشه بندی در بسیاری از کاربردهای پردازش متن مانند خلاصه سازی ,درک متن , ترجمه ماشینی و..استفاده می شود