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

# خلاصه سازی گزینشی متون فارسی مبتنی برخوشه بندی PSO

مهدی بازقندی[۱]، قمرناز تدین تبریزی[۲] , مجید وفایی جهان[۳] وعلی بازقندی [4]
Mehdi_Bazghandi@ymail.com[۱]
Tadayon@mshdiau.ac.ir [۲]
VafaeiJahan@mshdiau.ac.ir[۳]
Bazghandi@shahroodut.ac.ir [4]

## چکیده

**با افزایش روز افزون منابع متنی , هرروز بر گستره اطلاعات قابل دسترس برای کاربران افزوده می شود.به طوریکه دسترسی دقیق و درست به اطلاعات همواره یکی از مسائل اصلی در برخورد با انها بوده است.سیستم های خلاصه سازی خودکار می توانند نقش مهمی در پوشش قسمت های اصلی متن و برطرف کردن محدودیت های زمان داشته باشند.در این مقاله یک سیستم خلاصه ساز مبتنی برخوشه بندی جملات ارائه شده است. در این مقاله برای حل مسئله خوشه بندی چند روش معرفی شده است به طوریکه برای رسیدن به یک خوشه بندی با کیفیت از الگوریتم های هوش جمعی برای بهینه سازی روش های معرفی شده استفاده می شود.تمام روش های استفاده شده , ارتباط معنایی کلمات را با استفاده ازارتباطات انها در متن ,در نظر می گیرد . در نهایت بعد از خوشه بندی جملات با استفاده از معیارهای ذکر شده جملات مناسب و کاندید , از هر خوشه انتخاب می شوند.**
**برای ارزیابی یک مجموعه از خبرهای ورزشی فارسی انتخاب شده است. نتایج حاصل از ارزیابی روش پیشنهادي نشان می دهند که سیستم خلاصه ساز معرفی شده .خروجی های بهتری در مقایسه با روش های مشابه دیگر دارد.**

## کلمات کلیدی

**خلاصه سازی, خوشه بندی, شباهت معنایی , کارایی , الگوریتم PSO**

## ۱- مقدمه

با افزايش روزافزون منابع متني الكترونيكي در شبكه گسترده جهاني وب. نياز به دسترسي صحيح، سريع وآسان به اطلاعات بيشترمي شود. و نيروي انساني قابل توجهي كه مي تواند در كارهاي ديگر به كار گرفته شود، معطوف به تسهيل بخشيدن به دسترسي به اطلاعات است . بر اي يافتن مناسب و سريع سندهاي مورد نظر كاربر ، خواندن كامل متون بزرگ نامناسب است . سيستم خلاصه سازي اتوماتيك اين خلاء را پرمي كند .فرآيند فشرده سازي يك منبع به صورتي كه حاصل،درحجم كمتر حاوي اطلاعات مهم منبع اصلي باشد را خلاصه سازي گويند مزيت های عمده ي توليد خودكار خلاصه به وسيله ماشين عبارتند از:اندازه ي خلاصه قابل كنترل است، يعني ماشين ميتواند خلاصه را باتوجه به ميزان فشرده گی مورد نظر كاربر تهيه كند.محتواي آن قابل پيش بيني است.می توان مشخص كرد كه هر بخش از خلاصه مربوط به كدام بخش از متن اصلی است.به طور كلی دومدل اصلی برای خلاصه سازی متن وجود دارد. خلاصه گزينشی كه پركاربردترين نوع خلاصه سازی است كه در ان با توجه به معيارهاي آماري، شهودي و يا تركيبی از اين دو تهيه می شود.از آنجا كه در توليد اين دسته از خلاصه ها، جملات متن تغييرات نحوي و معنايی ندارند، می توان آن را نوعی گزينش جملات قلمداد كرد.و ديگری چكيده است كه تفسيری از متن اوليه است.در توليد چكيده ,خلاصه مفاهيم جملات متن اصلی به شكل كوتاه تر بازنويسی می شود.
فرايند خلاصه سازی به دو مدل تک سندی و چند سندی تقسم می شود.در مدل خلاصه ساز تک سندی , ورودی سيستم خلاصه ساز تنها يک سند می باشد[1] .در خلاصه ساز چند سندی , ورودی خلاصه ساز چندين سند می باشد.پيچيدگی خلاصه ساز تک سندی به علت پيوستگی موضوع در مقايسه با چند سندی , كمتر می باشد.دسته بندی های مختلفی برای روش های خلاصه سازی ارائه شده است.در مقالات زيادی تشابه دو جمله براساس كلمات كانديد ان (اسم ها و فعل ها) در WordNet در نظر گرفته شده است.در[2] از يک معيار مبتنی بر اشتراک كلمات , برای تعيين شباهت جملات استفاده می كند.مشكل اصلی اين روش ها در نظر نگرفتن معنای جملات می باشد. فقط به ظاهر كلمات توجه می كنند.درروشی ديگر از قانون احتمال شرطی بيزين برای احتمال حضور يک جمله در خلاصه استفاده می شود.[3].بسياری از مقالات از تكنيک خوشه بندی برای دسته بندی كردن جملات مشابه استفاده می كنند.در [4] با استفاده از يک روش مبتنی بر مركز جملات مشابه در يک خوشه قرار می گيرد و سپس با