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

که در خوشه بندی ان ها معیار شباهت استفاده شده,نقش مهمی در کیفیت خوشه ها خواهد داشت.در این بخش به چند مورد از کارهای انجام گرفته در این زمینه اشاره می کنیم.یکی از عمومی ترین الگوریتم های خوشه بندی , الگوریتم K-Meansمی باشد این الگوریتم با داشتن ساده گی و خطی بودن , کارایی مناسبی در برخورد با داده های حجیم دارد[1].در الگوریتم K-Means جواب نهایی وابستگی زیادی به انتخاب های اولیه مانند تعداد کلاسترها و مراکزنقاط دارند.ومتناظر با این انتخاب ها راه حل های محلی خواهیم داشت[2]. درمقاله [3] از الگوریتم PSO برای خوشه بندی داده ها استفاده می کند این الگوریتم مشکل پاسخ های بهینه الگوریتم K-Means را برطرف می کند و با جستجوهای بیشتر پاسخ گلوبال را بدست خواهد اورد . در همین مقاله ترکیب الگوریتم های K- Means و PSO در نظر گرفته شده است و جواب بهتری نسبت به هر کدام از الگوریتم ها دارد. در [4] از الگوریتم PSO برای خوشه بندی اسناد استفاده کرده است اما معیار استفاده شده فاصله اقلیدسی است .اما این معیار,فقط ظاهر کلمات موجود در متن را در نظرمی گیرد و به ارتباط معنایی ان ها توجهی نمی کند. درMead که یک سیستم خلاصه ساز مبتنی بر مرکز می باشد, با استفاده از خوشه بندی مبتنی بر مرکز,جملات مشابه یک متن را در یک خوشه قرار می دهد .[۵].در زبان فارسی در مقاله [۹] با استفاده از خوشه بندی سلسله مراتبی و K-Means جملات مشابه را دریک خوشه قرار می دهد..دراین مقاله سعی شده است ابتدا ارتباط معنایی کلمات با استفاده از روش های اماری تعیین شود. وسپس با استفاده از الگوریتم خوشه بندی PSO و بهبود یافته ان ,جملات یک متن را خوشه بندی کرد.

۲- روش پیشنهادی

در این مقاله ابتدا با استفاده از بردار های Context-Vector ارتباط معنایی کلمات تشخیص داده می شود.سپس الگوریتم PSO تشریح می شود. در ادامه ان الگوریتم خوشه بندی PSO پیشنهادی بیان می شود.ودر انتها روش های پیشنهادی با یکدیگرمقایسه می شوند.

۲-۱-ارتباط معنایی واژه ها

در ابتدا نیازداریم پیکره خود را مورد پردازش قرار دهیم و اسامی موجود در ان را به عنوان کلمات کاندید استخراج کنیم. در زبان انگلیسی می توان به ابزاری مانندTNT-Tager و SENTA اشاره کرد که اسامی یک جزئی یا دو جزئی را استخراج می کند. با توجه به اینکه در زبان فارسی با کمبود این ابزار ها و منابع رو به رو هستیم .در این مقاله سعی شده است اسامی موجود در پیکره از قبل به صورت دستی در یک فایل ذخیره شوند.هدف این بخش ان است که ارتباط و میزان شباهت تمامی اسامی استخراج شده , تعیین شود ودر ماتریس شباهت ذخیره شود. ما برای رسیدن به این هدف بردارهای Context Vector را معرفی می کنیم .برای هر کدام از واژه های اسمی موجود در متن یک بردار N بعدی تعیین می کنیم که شامل N کلمه با بیشترین ارتباط با واژه مورد نظر است. برای ای منظور ما باید بیشترین رخدادهای همزمان این واژه با سایر واژه ها را پیدا کنیم. احتمال شرطی متقارن (Symmetric Conditional Probability) روشی را معرفی می کند که می توان میزان ارتباط و همزمانی واژه ها را بدست اورد[۶] که معادله ان به صورت زیر است :

$$SCP(w_1, w_2) = \frac{p(w_1, w_1)}{p(w_1)p(w_2)} \qquad (۱)$$

که $p(w_1, w_1)$ احتمال رخداد همزمان دو واژه در یک همسایگی معین است. $p(w_2)$ و $p(w_1)$ به ترتیب احتمال رخداد واژه w_2 و w_1 به تنهایی را بیان می کنند.در پیاده سازی یک پنجره متنی به اندازه ۲۰ واژه لحاظ شده است .طبق اندازه پنجره, می توان واژه هایی که در همسایگی "باشگاه استقلال" امده را مشاهده کرد.(جدول(۱))

جدول(۱).رخداد همزمان واژه ها با "باشگاه استقلال"

مدیر برنامه‌هاي مجتبي جباري گفت: جباري امروز قصد آشتي با مسوولان باشگاه استقلال را داشت اما، صحبت‌هاي كادر فني اين تيم همه چيز را به هم زد. خشايار محسني در گفت‌وگو با خبرنگار ورزشي خبرگزاري دانشجويان ايران (ايسنا) گفت: زماني كه روز پنج‌شنبه به باشگاه استقلال رفتيم، برخورد بدي با جباري صورت گرفت. البته پس از اين جلسه، مسوولان باشگاه استقلال از اقدامشان اظهار پشيماني كردند و من نيز با جباري صحبت كردم. به همين دليل اين بازيكن قصد داشت امروز براي آشتي به باشگاه استقلال برود اما، صبح با من تماس گرفت و با اشاره به مصاحبه‌هاي حجازي و حاجيلو در مطبوعات، از تصميم خود منصرف شد.

بعد از استخراج بردارهای Context-Vector مربوط به اسامی می توان میزان مشابهت دو کلمه را در فضای برداری مقایسه کرد. و چنانچه دو بردار در فضای برداری بهم نزدیک باشند نشان می دهد که دو واژه به یکدیگر نزدیک می باشند و در یک متن مشابه بکار گرفته شده اند.و ارتباط ان ها تقریبا مشابهت است. در صورتی واژه هایی که در بردارContext-Vector یک واژه می ایند روابط متنوعی با ان خواهند داشت. اکنون برای بدست اوردن واژه هایی که برداره های ان ها بهم شبیه است به صورت زیر عمل می کنیم.در این مرحله میزان شباهت واژ ه های اسمی را می باییم.فرض کنیم که بردار زیر موجود است.

$$x_i = (x_{i,1}, x_{i,2} .... x_{i,p}) \qquad (2)$$

این بردار مربوط به واژه اسمی X می باشد.که میزان شباهت اطلاعاتی دو واژه $S_{i,j} = f(x_i, x_j)$ بصورت زیر محاسبه می شود[6] .

(3)

$$\inf osimBA(x_i, x_j) = \frac{A_{I,J}}{B_i * B_j + A_{i,j}}$$

$$A_{i,j} = \sum_{k=1}^{p}\sum_{l=1}^{p} X_{i,k} * X_{j,l} * SCP(w_{i,k}, w_{j,l})$$

$$\forall i, B_i = \sqrt{\sum_{k=1}^{p}\sum_{l=1}^{p} X_{i,k} * X_{j,l} * SCP(w_{i,k}, w_{j,l})}$$

که در این رابطه $X_{I,J}$ مربوط به وزن واژه $W_{I,J}$ است .[4] برای ۱۰۰ واژه اسمی مهمتر میزان تشابه انهارا به کمک تابع $f(x_i, x_j)$ حساب می کنیم. و در یک ماتریس ذخیره می کنیم.به جدول(۲) نگاه کنید.

جدول(۲):Context Word

لیگ برتر	مدرعامل ساپیا۰,۳۸	علی دایی ۰,۳۹	سرمربی ۰,۵۰	ساپیا ۰,۷۵
جباری	فتح اله زاده	تیم ملی ۰,۲۳	مدیر عامل ۰۲۸	استقلال ۰,۳۹
		۰۵۲		
قرارد	میلان ۰,۶۳	بارسلونا ۰,۷۲	الکلاسیکو ۰,۸۵	رونالدو ۰,۱۹
رئال مادرید	اسپانیا ۰,۵۶	گواردیولا ۰,۸۱	نیوکمپ ۰,۶۲	بارسلونا ۰,۴۹
سرمربی	اسیا ۰,۴۲	تیم ملی ۰,۵۱	ساپیا ۰,۵۴	علی دایی ۰,۶۲

با استفاده از رابطه شباهت (۳) (تشابه برداری) میزان شباهت دو واژه را دریک ماتریس ذخیره می کنیم این رابطه نشان می دهد که دو واژه تا چه اندازه از لحاظ معنایی به همدیگر مرتبط هستند.و جنانچه این اندازه از یک میزان استانه بیشتر باشد نشان می دهد که این دو واژه به یکدیگر مشابه هستند.و ماتریس شباهت را می سازد. این ماتریس برای یکبار ساخته می شود و ذخیره خواهد شد.جدول(۳).

جدول(۳):ماتریس شباهت

دروازه بان استقلال	سید مهدی رحمتی	اقای گل جهان	علی دایی	
۰,۲۱	۰,۳۲	۰,۶۵	۱	علی دایی
۰,۱۴	۰,۳۴	۱	۰,۶۵	اقای گل جهان
۰,۷۱	۱	۰,۶۵	۰,۳۲	سید مهدی رحمتی
۱	۰,۷۱	۰,۱۴	۰,۲۱	دروازه بان استقلال

اکنون با استفاده از ماتریس شباهت واژه ها ,شباهت میان دو جمله به صورت زیر حاصل می شود[6].

$$Sim(s_i, s_j) = \sum_{k=1}^{n}\sum_{i=1}^{m} \inf osimBA(x_i, x_j) \qquad (۴)$$

به دو جمله زیر دقت کنیم:

۱) علی دایی دروازه سید مهدی رحمتی را گشود.
۲) برترین گلزن جهان, دروازه بان استقلال را مغلوب کرد.

این دو جمله از لحاظ ظاهری هیچ شباهتی به یکدیگر ندارند.و فاصله کسینوسی میزان شباهت ان ها را صفر فرض می کند.در صورتی که دو جمله بالا از لحاظ معنایی یکسان است.و رابطه (۲) میزان شباهت ان هارا عددی مخالف صفر فرض می کند.و نشان می دهد این دوجمله با یکدیگر ارتباط دارند.

به ماتریس جمله- کلمه دو جمله بالادقت کنید:

جدول (۴) ماتریس جمله-کلمه

	علی دایی	سید مهدی رحمتی	برترین گلزن جهان	دروازه بان استقلال
S1	1	1	0	0
S2	0	0	1	1

طبق رابطه کسینوسی:

$$Co\sin e(s_1,s_2)=\frac{\sum_{i=1}^{n} S_i * S_j}{\sqrt{(S_i)^2}*\sqrt{(S_j)^2}} \qquad (۵)$$

میزان تشابه دو جمله S_2 و S_1 طبق رابطه بالا صفر می شود.درصورتی که طبق رابطه (۴) میزان تشابه ان ها برابر ۰,۵۲ بدست خواهد امد.بنابراین با این رابطه میزان تشابه دو جمله را حساب می کنیم و در یک ماتریس ذخیره می کنیم.ماتریسی شبیه زیر حاصل خواهد. شد.جدول (۵)

خلاصه این روش به این صورت است : برای یک متن ورودی ابتدا برای هر جمله اسامی موجود در تکتک جملات استخراج می شوند . وزن هر کدام از ان ها بدست می اوریم .(معیار TF-ISF). سپس با استفاده از ماتریس شباهت و رابطه (۴) می توان میزان شباهت جملات را محاسبه کرد.و ماتریس شباهت جملات را بدست اورد.

جدول(۵):ماتریس شباهت جملات

	S_1	S_2	S_3
S_1	۱	۰,۶۵	۰,۲۳
S_2	۰,۶۵	۱	۰,۳۴
S_3	۰,۲۳	۰,۳۴	۱

۲-۲-الگوریتم PSO

الگوريتم اجتماع ذرات يك الگوريتم بهينه سازي تقليدي از رفتارهاي جوامع جانوري در پردازش دانش جامعه است. اين الگوريتم از دو زمينه ريشه گرفته است. نخست زندگي مصنوعي (مانند دسته پرندگان، ماهي ها)و دوم محاسبات تكاملي [8].مبناي توسعه الگوريتم PSO اين است كه جواب هاي ممكن در يك مسأله بهينه سازي به صورت پرندگاني بدون حجم و خصوصيات كيفي در نظر گرفته مي شوند كه از آنان به عنوان ذرات ياد مي شود، اين پرندگان در يك فضاي n بعدي پرواز كرده و مسير حركت خود در فضاي جستجو را بر اساس تجارب گذشته خود و همسايگانشان تغيير مي دهند.در دسته اي متشكل از n جز ، موقعيت جزء iام تحت اثر يك بردار مكاني n بعدي مطابق معادله (۶) قرار دارد. اين جزء همچنين داراي يك بردار سرعت به صورت معادله (۷) مي باشد. بهترين موقعيت قبلي بدست آمده براي جز iام، با استفاده از معادله (۸) نمايش داده مي شود.

$$x_i=(x_{i,1},x_{i,2},,,,,,x_{i,n}) \qquad (6)$$

$$v_i=(v_{i,1},v_{i,2}.....v_{i,1}) \qquad (7)$$

$$p_i=(p_{i,1},p_{i,2},.....p_{i,n}) \qquad (۸)$$

در نهايت موقعيت جديد اجزاي دسته با استفاده از معادلات زير به دست مي آيد.

$$v_i(t+1)=\Omega v_i(t)+c_1r_1(p_1(t)-x_i(t))+c_2r_2(p_g(t)-x_i(t))$$
$$x_i(t+1)=x_i(t)+v_i(t+1) \qquad (9)$$

۲-۳-الگوریتم PSO باینری

.هر ذره در این الگوریتم یک مقدار باینری دارد. ومقدار ان با یک تابع برازندگی سنجیده می شود.اما به جای استفاده از معادله سرعت , موقعیت جدید ذره به صورت زیر حساب می شود.

$$P_{IJ}(t)>=\frac{1}{1+\exp(-v_{i,j}(t))}\rightarrow X_{ij}(t+1)=0 \qquad (10)$$
$$else\rightarrow X_{ij}(t+1)=1$$

۲-۴-خوشه بندی معنایی PSO

روش پیشنهای ما به اینصورت است که به جای استفاده از فاصله اقلیدسی دو بردار یا فاصله کسینوسی انها, از فاصله معنایی انها استفاده کنیم.ما در مراحل قبل ,میزان شباهت معنایی دو جمله را با استفاده از بردار های متنی بدست اوردیم.در این روش باید دو معیار مهم در نظر گرفته شوند.

۱)جملات باید طوری خوشه بندی شوند که جملات موجود در یک خوشه باید دارای بیشترین شباهت ممکن به یکدیگر باشند.(تابع درون خوشه ای Intra Clustering)

۲) جملات یک خوشه باید دارای فاصله زیادی با خوشه های دیگر داشته باشند.(تابع برون خوشه ای Inter Clustering)

بنابراین توابعی که بتواند این معیارها را ارضاء کنند به صورت زیر تعریف می شود:

تابع درون خوشه ای:(Intra Clutering)

$$p=(\sum_{p=1}^{k} c_p \sum_{s_i,s_j} sim(s_i,s_j) \rightarrow MAX \qquad (11)$$

تابع برون خوشه ای(Inter Clustering)

$$R=\sum_{P=1}^{K-1}\sum_{Q=P+1}^{}\sum_{s_i\in c_p}\sum_{s_j\in c_q} c_p c_q sim(s_i,s_j) \rightarrow MIN \quad (۱۲)$$

توابع ترکیبی

$$H = P*R \rightarrow MAX \qquad (۱۳)$$

$$E = w_2 * p + (1-w_2) * R \rightarrow MAX \qquad (14)$$

$$F = \frac{2}{\frac{1}{P}+\frac{1}{R}} = \frac{2*P*R}{P+R} \rightarrow MAX \qquad (۱۵)$$

تابع ترکیبی H ,تابع حاصل ضرب دو تابع P و R است. تابع E تابع ترکیب وزنی دوتابع است.وتابع F میانگین هارمونیک دو تابع می باشد.تابع E نشاندهنده میزان مشارکت دو تابع P و R می باشد به طوری که اگر w_2 برابر صفر باشد تابع E نشاندهنده تابع R خواهد بود و چنانچه یک انتخاب شود نشاندهنده P خواهد بود.و اگر ۰,۵ انتخاب شود, هر کدام از توابع(P,R) سهم یکسانی را خواهند داشت. در این توابع K تعداد خوشه ها ست. C_p تعدادجملات موجود در خوشه است . $sim(s_i,s_j)$ میزان شباهت ذخیره شده در ماتریس شباهت جمله (جدول(۵)) است.

۲-۴-۱-روش خوشه بندی پیشنهادی

الگوریتم های خوشه بندی خودکار, مجموعه ای از اشیاء را در تعدادی از خوشه قرار می دهند.به طوری که بهترین عدد برای تعداد خوشه ها توسط الگوریتم خوشه بندی مشخص می شود.در تهایت اشیاء موجود در هر خوشه باید بیشترین شباهت ممکن به یکدیگر را داشته باشند .در عین حال که با اشیاء دیگر خوشه ها فاصله داشته باشند. درروش خوشه بندی پیشنهادی , مجموعه ای از جملات $S=[s_1,s_2,s_3,....s_n]$ وجود دارند به طوری که باید در مجموعه ای از خوشه ها $C=\{C_1,C_2,....C_K\}$ بدون هیچ گونه همپوشانی ,قرار گیرند.تعداد خوشه ها ی بهینه (K) مشخص نمی باشد.این روش خوشه بندی دارای شرایط زیر می باشد:

۱) دو کلاستر مختلف, جمله مشترکی ندارند.

$$\forall p\neq q: p,q\in\{C_1,C_2,....C_K\}, C_P\cap C_q=\emptyset$$

2)هر جمله باید حتما در یک کلاستر قرار گیرد.

$$\bigcup C_P = D$$

3) هر کلاستر باید حداقل یک جمله را داشته باشد.

$$C_p\neq\emptyset \;\; \forall p\in\{1,2,3,...k\}$$

در الگوریتم PSO برای اینکه بتوانیم از روش خوشه بندی پیشنهادی استفاده کنیم, باید ساختار هر ذره طوری تعریف شود که سه شرط بالا را ارضا کند. .در این ساختار تعریفی , طول هر ذره برابر تعداد جمله های متن تعریف می شود. به طوری که برای هر جمله باید عددی از مجموعه اعداد {1,2,3,…K} انتخاب شود.(شرط دوم).در این ساختار باید از تمام K عدد موجود استفاده کرد.(شرط سوم).هر جمله فقط یک عدد منحصر به فرد دارد.(شرط اول). این عدد در طول ذره منحصر به فرد نمی باشد.بنابراین به طور خلاصه , اعداد موجود در هر ذره (با توجه به تعداد خوشه ها) باید طور انتخاب شوند که از همه اعداد {1,2,3,…K} استفاده شود.در طول ذره ممکن است یک عدد که نشان دهنده شماره کلاستر می باشد ,تکرار شود. اما باید از همه این K عدد استفاده شود.دراین مقاله ساختار هر ذره مانند شکل(۱) تعریف می شود.

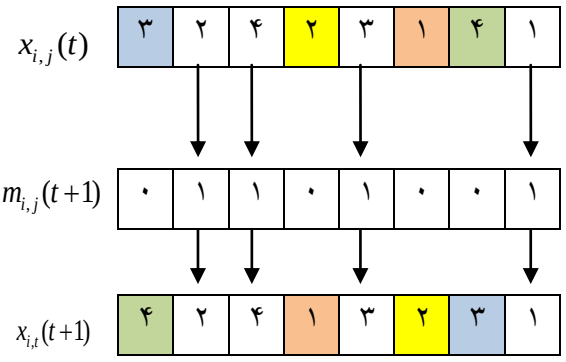
۴	۴	۳	۱	۲	۲	۳	۱	۱

شکل(۱) ساختار یک ذره

این ذره نشان می دهد که جملات اول دوم و ششم در کلاستر اول, جملات سوم و هفتم در کلاستر سوم, جملات چهارم و پنجم درکلاستر دوم و جملات هشتم ونهم در کلاستر چهارم می باشد. تابع برازنده گی مورد استفاده درالگوریتم , یکی از توابع درون خوشه ای , برون خوشه ای و توابع ترکیبی بخش 2-4 می باشد. مسئله ای که در این الگوریتم باید به ان اشاره کرد این است که ذرات تولیدی نباید مقادیری خارج از سه شرط ذکر شده تولید کنند.اما این مسئله در الگوریتم PSO اجتناب ناپذیر است.که ناشی از وجود بردارهای سرعت ذرات می باشد.که ممکن است مقادیر ذرات تولیدی خارج از محدودیت های ذکر شده (حتی اعشاری) شوند.که برای حل این مسئله می توان از الگوریتم ژنتیک بهره برد و برداری به نام بردار جهش به صورت زیر استفاده کرد.

$$m_{i,j}(t+1)=\begin{cases}1, if Rand_j \le sigm(v_{i,j}(t+1)\\ 0, otherwise\end{cases} \quad (16)$$

بردار $m_{i,j}(t+1)$ با تاثیرگذاشتن روی $x_{i,j}(t)$, ان را به بردار $x_{i,t}(t+1)$ تبدیل می کند.چنانچه j امین عنصر از بردار $m_{i,j}(t+1)$ برابر یک باشد.عنصر j ام بردار $x_{i,j}(t)$ بدون تغییر به j امین عنصر بردار $x_{i,t}(t+1)$ منتقل می شود.ودر صورتی که j امین عنصر بردار $m_{i,j}(t+1)$ صفر باشد.عنصرj ام بردار $m_{i,j}(t+1)$ جهش خواهد داشت. این جهش ناشی از تاثیر عملگر معکوس بروی ان می باشدشکل ۲. شکل گیری بردار $x_{i,t}(t+1)$ با استفاده از $m_{i,j}(t+1)$ رانشان می دهد.



شکل(۲)عملکرد بردار جهش

۳-تعداد خوشه های بهینه

شاخص های زیادی برای پیدا کردن تعداد خوشه های بهینه معرفی شده است که اغلب ان ها از دو پارامتر فشرده گی و تفکیک استفاده می کنند.که ما از تابع زیر استفاده می کنیم[7]:

$$V_{ram}(K)=\frac{comp(k)}{separ(k)} \quad (۱۷)$$

هنگامی که تعداد خوشه ها زیاد می شود.میزان فشرده گی خوشه ها کاهش می یابد.و مقدار $V_{ram}(K)$ کاهش میابد.در نتیجه $V_{ram}(K+1)\le V_{ram}(K)$..ما برای رسیدن به تعداد خوشه های بهینه از روش سعی وتکرار استفاده می کنیم به طوریکه هدف مینیم کردن تعداد خوشه هلیی است که در رابطه زیر صدق کند:

$$H=\frac{V_{ram}(K)-V_{ram}(K+1)}{V_{ram}(K)}\le \varepsilon \quad (۱۸)$$

که این رابطه تعداد خوشه های بهینه را تعیین می کند و ε یک استانه می باشد.

۴-نتایج

ما یک مجموعه خبر ورزشی مربوط به خبرگزاری دانشجویان ایران (ایسنا) را به عنوان پیکره متنی خود انتخاب کرده ایم .و تمامی ازمایشات و ارزیابی ها را روی این مجموعه داده انجام داده ایم.این مجموعه شامل ۸ مجموعه خبر ورزشی می باشد.دو روش ارزیابی برای سیستم های بازیابی اطلاعات وجود دارد ارزیابی مستقیم و ارزیابی معطوف به کاربرد. در ارزیابی مستقیم , نتیجه سیستم با نتیجه کار چند انسان بصورت مستقیم ارزیابی می شود. با توجه به اینکه این مقاله به خوشه بندی جملات فارسی پرداخته است و ابزاری برای ارزیابی ان وجود ندارد. از ارزیابی مستقیم استفاده خواهیم کرد.این روش ارزیابی مشابه روش استفاده شده در [5] می باشد.دو معیار مهم در خوشه بندی جملات عبارتند ازسودمندی نسبی درون خوشه ای (CBRU) واشتراک اطلاعات بین جمله ای(CSIS).در حالت ایده ال خوشه بندی مناسب است کهCBRU را که بیانگر میزان ارتباط جملات موجود در یک خوشه است, ماکزیمم کندو معیارCSIS را کاهش دهد.در این ارزیابی ها از چند داور خواسته می شود که به جملات موجود درتمام کلاسترها نمره دهند و سپس کارایی کلاستر مورد نظر بدست خواهد امد..(تمام

روش ها به ازای متن واحدی سنجیده می شود.) به شکل(۳) توجه کنید:

جدول (5):متن خبر:دیدار دکتر احمدی نژاد از اردوی تیم ملی

[1] حضور رئیس جمهور اسلامی ایران در تمرینات تیم ملی فوتبال، فضای خاصی را به اردوی تیم ملی بخشید.
[2] به گزارش خبرگزاری دانشجویان ایران (ایسنا)، به محض ورود رئیس جمهور بر سر تمرینات تیم ملی از درب ورودی ورزشگاه آزادی، تماشاگران تمرینات تیم ملی با شعار "انرژی هسته ای حق مسلم ماست" و "احمدی نژاد دوستت داریم" به تشویق او پرداختند.
[3] احمدی نژاد سپس با برانکو به گفت و گو پرداخت و نفس تلاش بازیکنان تیم ملی را ارزشمند و یک پیروزی خواند.
[4]رئیس جمهور در ادامه با تک تک بازیکنان خوش و بش کرد و ابراهیم میرزاپور به گفت و گو و درد دل با رئیس جمهور پرداخت.
[5] رییس جمهور در ادامه با تك تك بازیکنان خوش و بش کرد و ابراهیم میرزاپور به گفتوگو و درد دل با رییس جمهور پرداخت.
[6] گفتني است دكتر احمدي نژاد پيش از سخنانش، دقايقي را با ملي‌پوشان تمرين كرد و به زدن ضربات پنالتي پرداخت. رييس جمهور ۴ ضربه‌ي پنالتي به ابراهيم ميرزاپور دروازه‌بان تيم ملي زد كه ۳ تاي آن تبديل به گل شد و يكي از آنها را ميرزاپور دفع كرد.
[7]رئیس جمهور همچنین با حضور در رختکن تیم ملی و با پوشیدن گرمکن تیم ملی به زمین بازگشت.
[8] دکتر احمدی نژاد هنگام ورود به زمین چمن به گرم کردن خود

جدول (6).جواب همگرا در روش P

خوشه بندی	gbest	Iteration
(1,1,2,3,4,1,2,3,2)	2.0330	1
(1,2,2,2,2,2,3,3,4)	2.5123	20
(1,1,1,1,3,3,2,2,1)	2.7901	40
(1,1,1,1,3,3,2,2,1)	2.7901	50
(1,1,1,1,3,3,2,2,1)	3.2301	100
(1,1,2,3,3,1,1,2,3)	3.4578	120
(1,1,3,2,2,3,1,2,2)	3.4897	150
(1,1,3,2,2,3,1,2,2)	3.4897	180
(1,1,1,1,2,3,3,2,2)	4.7389	200
(1,1,1,1,2,3,3,2,2)	4.7389	220
(1,1,1,1,2,3,3,2,2)	4.7389	250

در جدول(6) همانطور که ملاحظه می شود پاسخ های گلوبال در تکرار۲۰۰ به به جواب نهایی همگرا شده است.(K=4) در همین روش نیاز است تعداد خوشه های بهینه مشخص شود.جدول.۵ با چهار خوشه در نظر گرفته شده است. تعداد خوشه های بهینه در جدول (7) تعیین شده است.(استانه 0.002).

جدول(۷):تعداد خوشه های بهینه در دروش P

K	1	2	3	4	5
H	0.007	0.006	0.003	0.002	0.001

اکنون خوشه بندی بدست امده از روش قبل باید ارزیابی شود.به اینصورت کند که تعدادی داور به تمام جملا ت موجود در یک کلاستر امتیاز می دهند این امتیازات(۰ تا ۱۰) باید طوری به جملات نسبت داده شود که تعیین کند جمله مورد نظر تا چه اندازه به موضوع عمومی خوشه نزدیک است.به جدول(۸) توجه کنید.(خوشه بندی بدست امده از مرحله قبل)

جدول(۸):ارزیابی یک کلاستر

داورها	داور اول	داور دوم	داور سوم	مجموع
داور اول	۱	۰,۴۳	۰,۳۵	۰,۵۹
داور دوم	۰,۴۳	۱	۰,۲۹	۰,۵۷
داور سوم	۰,۳۵	۰,۲۹	۱	۰,۵۴
				۰,۵۶

با توجه به اینکه از هر خوشه دوجمله با امتیازبالا (متناسب با نرخ فشرده گی) توسط داورها انتخاب می شود .کارایی کلاستر اول توسط روش P برابر ۰,۵۶ است .بقیه کلاسترها ی این روش توسط داورها ارزیابی می شوند.سپس میانگین انها, کارایی کل خوشه بندی را نشان خواهند داد. در تمامی این روش ها پارامترهای الگوریتم PSO به اینصورت مقداردهی شدند:جمعیت اولیه ذرات برابر ۴۰ است.(N=40) .تعداد تکرار ۲۵۰ (t=250) و پارامترهای شناختی و اجتماعی $c_1 = 2.5, c_2 = 4 - c_1$ و وزن اینرسی برابرw =۰,۵ است.تمام نتایج ثبت شده, حاصل اجرای میانگین ۱۰ بار اجرای هر روش می باشد در .تمام ارزیابی ها در تمام این روش ها با تعدا د خوشه های بهینه انجام می گیرد.بنابراین ابتدا باید تعداد خوشه های بهینه هر روش را محاسبه کرد. و در روش E فاکتور $w_2 = 0.4$ در نظر گرفته شده است.انتخاب این مقدار برای w_2 با استفاده ازارزیابی تاثیر ان روی مسئله کارایی صورت گرفته است[10]. برای ارزیابی نهایی از هشت مجموعه خبری استفاده شده است.به طوریکه نتایج ثبت شده در جدول (۹)

و(۱۰) کارایی روش های خوشه بندی را روی مجموعه های خبری(دیدار احمدی نژاد از اردوی تیم ملی و قهرمانی سایپا) نشان می دهد.

جدول (۹):مقایسه روش های خوشه بندی روی مجموعه خبراول

	کارایی خوشه بندی	P	R	H	E	F
P	0.515	X	-0.040	+0.006	+0.013	+0.021
R	0.495	+0.038	X	+0.044	+0.051	+0.058
H	0.518	-0.006	-0.046	X	+0.008	+0.025
E	0.522	-0.013	-0.600	-0.007	X	+0.008
F	0.526	-0.021	-0.062	-0.015	-0.007	X

در جدول(۹) برای تعیین کردن میزان بهبودی روش ها در مقایسه با یک روش خاص ,از فرمول زیر استفاده می شود:

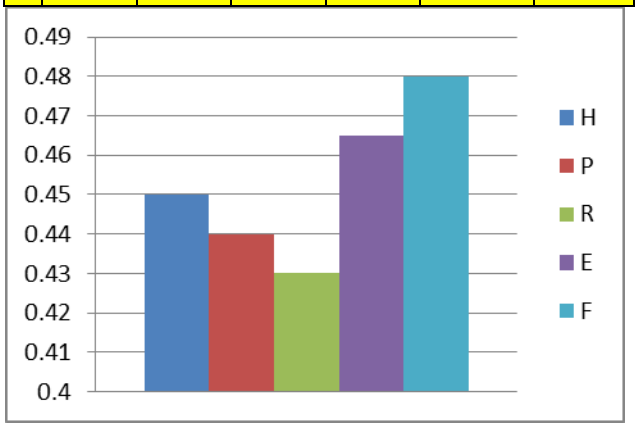
$$\frac{OurMethod - OtherMethod}{OtherMethod} * 100 \qquad (19)$$

علامت مثبت نشاندهنده میزان بهبودی یک روش نسبت به روش خاص دیگر را نشان می دهد و علامت منفی نشان دهنده عکس ان است.به طور مثال کارایی روش H برابر ۰,۵۱۸ است و روش E وF در مقایسه با روش H , به ترتیب به اندازه ۰,۸ و ۲,۵ درصد عملکرد بهتری نسبت به ان دارند. جدول زیر عملکرد روش های مختلف خوشه بندی روی مجموعه خبری دو(قهرمانی سایپا را نشان می دهد).

جدول (۱۰):مقایسه روش های خوشه بندی روی مجموعه خبردوم

با توجه به جدول (۱۰) همانطورکه ملاحظه می شود روش خوشه بندی F عملکرد بهتری در مقایسه با روش های دیگر دارد.با اجرای تمام روش ها روی هشت مجموعه خبری نتایج نشان می دهد روش خوشه بندی میانگین هارمونیک (F) ,کارایی بهتری در مقایسه با روش های دیگر دارد.درمقابل روش های P و R عملکرد ضعیفی نسبت به بقیه خواهند داشت.(میانگین روی هشت مجموعه خبر)

	کارایی خوشه بندی	P	R	H	E	F
P	۰,۴۲۱	X	+0.119	+0.276	+0.266	+0.211
R	۰,۴۷۸	-0.135	X	+0.078	+0.096	+0.104
H	۰,۵۱۹	-0.232	-0.085	X	+0.018	+0.028
E	0.۵۲۹	-0.256	-0.628	-0.019	X	+0.009
F	۰,۵۳۴	-0.268	-0.117	-0.028	-0.009	X



شکل(۳): مقایسه روش های خوشه بندی

۵-نتیجه گیری

یک مسئله مهم در خلاصه سازی متون, استخراج موضوعات و جملاتی است که تا حدود زیادی بتوانند مفهوم متن اصلی را برسانند.در سال های اخیر مقالات زیادی به این موضوع پرداخته اند.و اغلب ان ها برای پیدا کردن شباهت میان متن ها و جملات از معیار فاصله کسینوسی و اقلیدسی استفاده کرده اند.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.ما برای حل این مشکل در این مقاله با استفاده از بردارهای Context-Vector ارتباط معنایی واژه ها را یافتیم و سپس با تعریف معیارها و توابعی که سعی در ارضاء ویژگی های مسئله خوشه بندی داشتند, توانسیم توابع مناسب را استخراج کنیم.وبرای بهینه شدن هر کدام از این توابع الگوریتم PSO رابکار برده ایم . ما در این مقاله چند تابع متناظر با ویژگی های خوشه بندی تعریف نمودیم. روش R بالا بودن میزان فشرده گی خوشه ها را درنظر می گیرد.درروش P بالا بودن

تعداد خوشه ها اهمیت می یابد. در بین روش های ترکیبی با توجه به نتایج حاصل ار ارزیابی روش F (تابع میانگین هارمونیک) مناسبترین روش به لحاظ کارایی معرفی شد.

٦-مراجع

[1] Ching-Yi Cheo , Fun Ye.”Particle Swarm Optimization Algorithm and Its Application”. lntemationai Conference on Networking, Sensing Control. Taiwi. Taiwan. March 21-23, 2004

[2] Zalik, Krista Rizman . “An Efficient k`-means Clustering Algorithm.”Pattern Reconition Letters, Vol. 29, I.9. Pag. 1385-1391. Elsevier 07/2008. to Clustering Analysis.

[3] R. J. Kuo • M. J. Wang • T. W. Huang .” An application of particle swarm optimization algorithm to clustering analysis “ .Springer-Verlag 2009.

[4] Ramanji Killani, K. Srinivasa Rao.“Effective Document Clustering with Particle Swarm Optimization”. Springer-Verlag Berlin Heidelberg 2010.

[5] D. R. Radev, H. Jing, M. Stys, and D. Tam (2004), Centroid-based summarization of multiple documents. Information Processing and Management, 2004.

[6] Dias, G. and Alves, E. (2005). Unsupervised Topic Segmentation Based on Word Co-occurrence and Multi-Word Units for Text Summarization. In Proceedings of the ELECTRA Workshop associated to 28th Annual International ACM SIGIR Conference, Salvador, Brazil, pp. 41-48.

[7] R.M.Aliguliyev.”Automation Documents Summarization By Sentence Extraction:.

[8] Eberhart R, Kennedy J (1995) A new optimizer using particle swarm theory. In: Proceedings of the sixth international symposium on micro machine and human science, pp 39–43

[۹] مشکی محسن, انالویی مرتضی " بازیابی خبرهاي مرتبط پیشین براي تولید خلاصه هاي پیشینه-خبر" اولین کنفرانس برق , دانشگاه علم و صنعت 1388

[۱۰] بازقندی ,مهدی, تدین تبریزی, قمرناز.وفایی جهان مجید."خلاصه سازی متون فارسی مبتنی بر خوشه بندی PSO ".اولین کنفرانس بین المللی پردازش خط و زبان فارسی.دانشگاه سمنان,۱۳۹۱.