یازدهمین کنگره ملی سراسری
فناوریهای نوین در حوزه توسعه پایدار ایران
11th National Congress of
the New Technologies in Sustainable Development of Iran
senaconf.ir

بهینه سازی یک شبکه اجتماعی آنلاین مبتنی بر تکنیک های داده کاوی

سید حامد میرمحمدی، منیره عیوضی لوشکانی

آزاد، تهران setaremonir@yahoo.com

فاران، تهران setaremonir@yahoo.com

چکیده

هدف پژوهش حاضر توسعه یک سیستم پیشنهاددهنده‌ی موثر و دقیق در شبکه‌های اجتماعی است که با ارائه مدلی جدید به خوشه‌بندی و تطبیق کاربران در شبکه های اجتماعی به ویژه شبکه های همسریابی، کمک می کند. جامعه آماری تحقیق، کاربران شبکه همسریابی آنلاین RSVP استرالیا می‌باشند. طراحی پژوهش حاضر سه فاز اصلی را پوشش می‌دهد که به ترتیب شامل پیش پردازش و پروفایل بندی اطلاعات کاربران، خوشه‌بندی و نهایتاً پیشنهاددهی کلاسترهای تطبیق یافته می‌باشد. در این تحقیق از الگوریتم میانگین K برای خوشه بندی و از الگوریتم درخت تصمیم‌گیری J48 جهت تطبیق موثر کاربران استفاده شده است. نتایج تحقیق نشان می‌دهد که روش پیشنهادی، نسبت به سیستم‌های سنتی تطبیق که کاربران را دو به دو به هم پیشنهاد می‌دهد، تطبیق موثر و قابل قبول‌تری ایجاد نموده است.

واژه‌های کلیدی: شبکه اجتماعی آنلاین، شبکه های همسریابی، الگوریتم میانگین K

مقدمه

تعداد شبکه‌های اجتماعی و اعضای آن‌ها به سرعت رو به افزایش است. این فرآیند یافتن دوستان جدید را روی شبکه اجتماعی بسیار چالش برانگیز ساخته است و نیاز به یک سیستم تطبیق اجتماعی را مطرح ساخته است که برای کاربران لیستی از دوستان پیشنهادی را ایجاد نماید. یک سیستم تطبیق اجتماعی می‌تواند به عنوان یک سیستم پیشنهاددهنده شناخته شود که به افراد، افراد دیگر را به جای پیشنهاد دادن آیتم‌ها برای آن‌ها پیشنهاد می‌دهد(۱).

یک نوع جدید از سیستم پیشنهاددهی، سیستم تطبیق اجتماعی، در چند سال اخیر به دلیل شهرت شبکه‌های اجتماعی ارائه شده است. چنین سیستم‌های تطبیق اجتماعی افراد را به یکدیگر پیشنهاد می‌دهد که این اختلاف اصلی آن با سیستم‌های پیشنهاددهی استاندارد است که محصولات را به افراد پیشنهاد می‌دهند. با این حال پیشنهاددهی افراد به یکدیگر پیچیده تر و چالش برانگیزتر است زیرا جنبه‌های زیاد را شامل می‌شود مانند حریم شخصی، انواع مختلف دیتاها، و ماهیت روابط. در نتیجه استفاده از تکنیک‌های سیستم‌های پیشنهاددهنده‌ی استاندارد برای تطبیق افراد در شبکه‌های اجتماعی منجر به پیشنهادات با دقت پایین می‌شود(۱). مسئله عمده دیگر عبارت از پیچیدگی سیستم تطبیق اجتماعی با استفاده از تکنیک‌های سیستم‌های پیشنهاددهی نرمال است. گرچه یک تعداد از شبکه‌های اجتماعی از قبل به طور گسترده کاربرد داشته اند دقت و راندمان آنها بنا به علل متعدد از دیدگاه‌های مختلف نیاز به بهبود و توسعه دارد(۲).

تحقیق فعلی با هدف بهبود فرآیند تطبیق در شبکه‌های اجتماعی همسریابی با به کارگیری تکنیک‌های خوشه‌بندی است(۳). در این زمینه تحقیقاتی نیز انجام شده است. هارپ، سن و فرانکفسکی در ۲۰۰۷ الگوریتمی طراحی کردند که به صورت اتوماتیک کاربران را در شبکه‌های اجتماعی گروه بندی کرده و گروه‌های حاصله را توصیف می‌کند. الگوریتم آن‌ها نیازی به گروه‌های از پیش ایجاد شده ندارد در عوض بر اساس خوشه‌بندی فعالیت بالانس شده است که هر دوی فعالیت کاربر و نیز علاقه مندی کاربر را در نظر می گیرد(۴). یک مقاله‌ی جالب توسط شرکت سان در ۲۰۰۵ از یک مدل تانسور برای بهبود شخصی سازی جستجوی وب استفاده کرده است. دیتای جمع‌آوری شده از یک موتور جستجوگر به وسیله یک مدل تانسور ۳ مرتبه ای ارائه شده است (۵).

پاول کری وسکی و همکارانش در ۲۰۰۷ مدلی ارائه دادند که نشان می داد شبیه سازی شبکه های اجتماعی با درجه توزیع یکسان می تواند رفتارهای خوشه بندی متفاوتی به دنبال داشته باشد(۶). در مقاله ای دیگر به نقش سیستم های توصیه گر و روش های خوشه بندی به منظور کمک به مشتریان فروشگاه های اینترنتی جهت تصمیم‌گیری بهتر در خرید کالا اشاره شده است(۷). در تحقیقی دیگر ضمن بیان رویکردهای پیشین به حوزه‌ی استفاده از داده کاوی در شبکه های اجتماعی، تلاش شده است تا مدل جدیدی با نگرش پویا و تاکید برانتخاب بهترین درخت های تصمیم ممکن متناسب با بانک داده ها ارائه شود(۸).

در جایی دیگر به مفاهیم داده کاوی و انواع تکنیک‌های داده کاوی در شبکه های اجتماعی پرداخته شده است و آن‌ها را در حوزه نظارت شده، نیمه نظارت شده و نظارت نشده بررسی می کند و در انتها به عنوان یک پیشنهاد تکنیک TRCM را جهت تشخیص وقایع قابل پیش بینی و غیرقابل پیش بینی ارائه می‌دهد(۹). در مقاله‌ای تحت عنوان "تحلیل رفتار کاربران در شبکه های اجتماعی با استفاده از داده کاوی" به بررسی قوانین پایه برای کشف اجتماعات پرداخته شده و در رابطه با برخی پیاده سازی‌های صورت گرفته در شبکه‌های اجتماعی مختلف از جمله یوتیوب در زمینه کشف اجتماع صحبت شده است(۱۰).

این تحقیق ارائه دهنده‌ی یک سیستم پیشنهاددهی مرکب و نوین است که از تکنیک خوشه‌بندی استفاده می‌کند. که از هر دو نوع دیتای ضمنی و صریح برای دستیابی به دقت بالاتر در پیشنهادات با پیچیدگی محاسباتی کمتر استفاده می‌کند. همچنین این طراحی به شبکه‌های اجتماعی در غلبه بر برخی از مشکلات مانند یک شروع دیرهنگام و یا پراکندگی کمک می‌کند.

از آنجا که خوشه‌بندی یک جزء اصلی در این تحقیق است، از الگوریتم میانگین k برای گروه‌بندی افراد و از دو الگوریتم خوشه‌بندی بر مبنای محدویت‌های افزایشی برای بهبود آن در شبکه‌های اجتماعی استفاده شده است. این الگوریتم‌ها برای غلبه بر نواقص در الگوریتم‌های خوشه‌بندی موجود و برای درنظر گرفتن نیازهای خاص افراد در شبکه‌های اجتماعی طراحی شده اند.

روش تحقیق

این تحقیق از نظر هدف توصیفی از نوع پیمایشی است. از نظر مکانی در طبقه تحقیق های اسنادی(کتابخانه ای) و میدانی(محیط طبیعی) قرار می‌گیرد. همچنین افق زمانی این تحقیق به صورت مقطعی می باشد. به منظور گردآوری اطلاعات نیز از پایگاه داده سایت همسریابی استفاده شده است. اطلاعات به کمک داده کاوی مورد تجزیه و تحلیل قرار گرفته اند.

این تحقیق یک روش جدید برای تطبیق کاربران در شبکه‌های اجتماعی با استفاده از داده‌کاوی ارائه داده است. این روش هم دیتای صریح(اطلاعاتی که کاربران در مورد خود ایجاد می‌کنند) و هم دیتاهای ضمنی(فعالیت‌های کاربران در شبکه‌های اجتماعی) را برای بهبود فرآیند تطبیق درنظر می گیرد. دیتای صریح برای کلاستر کردن کاربران مرد و زن در گروه‌های همگن به کار می‌رود. سپس کلاسترها با استفاده از دیتای ضمنی، کاربران را به یکدیگر پیشنهاد می دهد. طراحی تحقیقاتی این روش که در این

جا توصیف شده است سه فاز عمده را پوشش می‌دهد: پیش پردازش و پروفایل بندی اطلاعات کاربران، خوشه‌بندی کاربران و پیشنهاددهی پارتنر(شریک)های مطابقت یافته.

فاز۱: پیش پردازش و پروفایل بندی

وظیفه اصلی در فاز ۱ عبارت از پیش پردازش دیتای بدست آمده از یک شبکه اجتماعی، و آماده سازی برای فاز بعدی است. دیتای راجع به کاربران شبکه‌های اجتماعی می‌تواند از دو منبع بدست آید: به صورت صریح از روی پروفایل کاربران و به صورت ضمنی از روی ردیابی رفتار کاربران در شبکه. نمونه‌های دیتای ضمنی عبارتند از: ارتباط با سایر کاربران، مانند ارسال دعوت نامه‌ها، تماشای پروفایل کاربران، و غیره.

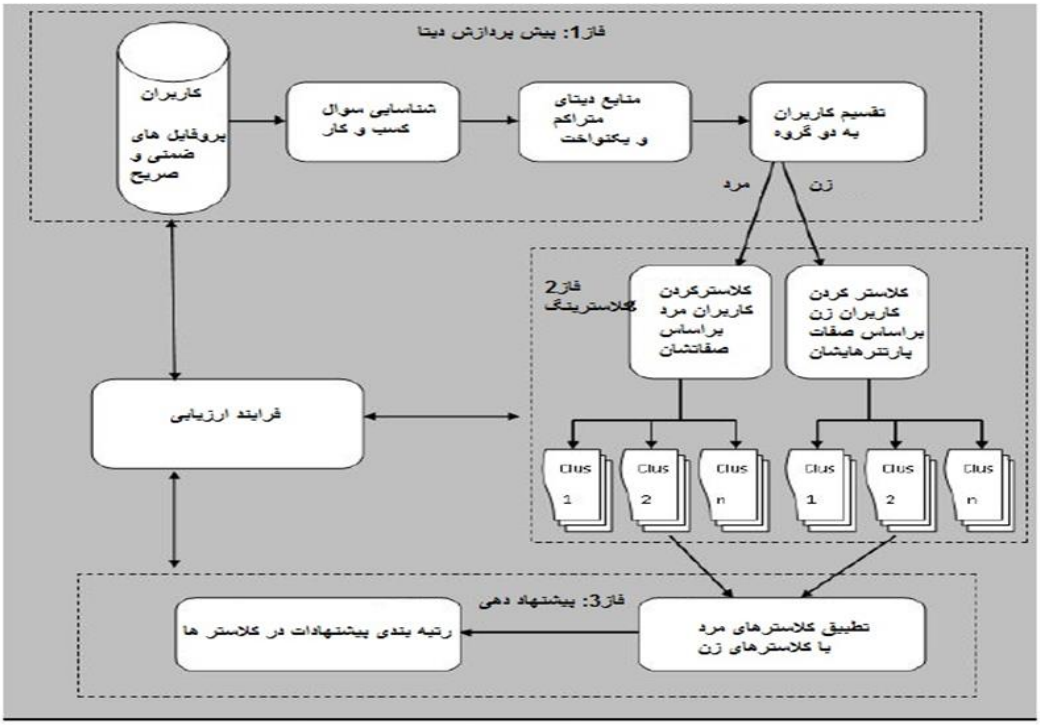
فاز ۲: خوشه‌بندی

هدف از فاز دو عبارت از کلاستر کردن کاربران مرد و کاربران زن به صورت جداگانه بر اساس تشابهات آن‌ها است. در این فاز، از الگوریتم میانگین k برای این منظور استفاده می کنیم و همچنین از دو الگوریتم افزایشی محدودیت‌های شبکه‌های اجتماعی برای کلاستر کردن کاربران، بهره می بریم. الگوریتم‌های جدید بر مسئله‌ی بُعد بندی(ابعاد) و پراکندگی در دیتاهای شبکه‌های اجتماعی غلبه می‌کنند.

فاز۳: پیشنهاد دهی

فاز ۳ با هدف پیشنهاد دادن بهترین پارامترها به کاربران بر اساس سازگاری آنها است. فرآیند پیشنهاددهی کاربران به دو بخش تقسیم می‌شود: تطبیق کلاسترها و رتبه‌بندی پیشنهادات. خروجی خوشه‌بندی عبارت از مجموعه ای از کلاسترهای مرد و مجموعه ای از کلاسترهای زن است. تطبیق کلاستر به تخصیص دادن یک کلاستر جنسیت هم تراز و مطابق با سازگارترین کلاستر جنسیت دیگر و باتوجه به تشابه کل کلاسترها تمایل دارد. الگوریتم توسعه یافته، تشابهات میان کلاسترها را بر اساس ویژگی‌های کلاسترها شناسایی می‌کند که دراین فرآیند به کار می‌رود

رتبه‌بندی با استفاده از الگوریتم j48 ، از الگوریتم های ساخت درخت تصمیم‌گیری انجام خواهد شد.بدین صورت که هر کاربر مرد در خوشه مشخص، کاربران زن در خوشه تطبیق یافته را بر اساس بیشترین مقادیر مشترکی که در صفاتشان دارند رتبه‌بندی می کند.



شکل۱: مدل پژوهش

در این تحقیق ۲۱ صفت برای کاربران در نظر گرفته شده است که کاربران می توانند این صفات را در پروفایل خویش تکمیل نمایند.

جدول۱: صفات کاربران و مقادیرمجزای آنها

ردیف	نام متغیر		مقادیر ممکن		تعداد مقادیر
۱	جنسیت	Gender (GEN)	Male (M)	مرد	۲
			Fmale (F)	زن	
۲	سطح شغلی	Skill (SKI)	Beginner	مبتدی	۳
			Technician	تکنسین	
			High	مهارت بالا	
۳	گروه سنی	Age Category (AGE)	Teenager	نوجوان	۴
			Young	جوان	
			Middle	میانسال	
			Old	پیر	

۴	وزن	Weight (WEI)	Thin	لاغر	۳
			Middle	متوسط	
			Fat	چاق	
۵	شخصیت	Character (CHA)	Introverted	درون گرا	۲
			Extroverted	برون گرا	
۶	صنعت و پیشه	Craft (CRA)	Unemployed	بیکار	۶
			Student	دانشجو	
			Worker	کارگر	
			Employee	کارمند	
			Self	آزاد	
			Oder	سایر	

ردیف	نام متغیر		مقادیر ممکن		تعداد مقادیر
۷	وضعیت تاهل	Marital Status (STA)	Single	مجرد	۴
			Married	متاهل	
			Divorced	مطلقه	
			Widow	بیوه	
۸	قد	Length (LEN)	Short	کوتاه	۳
			Normal	متوسط	
			Tall	بلند	
۹	تحصیلات	Education (EDU)	Basic	بی سواد	۳
			Elementary	ابتدایی	
			Collegiate	دانشگاهی	
۱۰	سیگار کشیدن	Smoking (SMO)	NO	بلی	۲
			YES	خیر	

۱۱	تمایل فرزند	Desire for having Chid (DCHI)	NO	بله	۲
			YES	خیر	
۱۲	رنگ چشم	Eye Color (EYE)	Black	مشکی	۶
			Brown	قهوه ای	
			Gray	طوسی	
			Blue	آبی	
			Green	سبز	
			Other	سایر	
۱۳	سطح اعتقادات	Believe (BEL)	H Religious	مذهبی مقید	۳
			M Religious	مذهبی معمولی	
			S Religious	غیرمذهبی	

ردیف	نام متغیر		مقادیر ممکن		تعداد مقادیر
۱۴	برج تولد	Birth Month (BIR)	January	ژانویه	۱۲
			February	فوریه	
			March	مارس	
			April	آوریل	
			May	می	
			June	جون	
			July	جولای	
			August	اوت	
			September	سپتامبر	
			October	اکتبر	
			November	نوامبر	
			December	دسامبر	
۱۵	رژیم غذایی	Diet (DIE)	NO Diet	بله	۲

			YES Diet	خیر	
۱۶	دین	Religion (REL)	Islam	اسلام	۴
			Christ	مسیح	
			Jewish	یهود	
			Other	سایر	
۱۷	رنگ مو	Hair Color (HAI)	Black	مشکی	۳
			Brown	قهوه ای	
			Gold	طلایی	
۱۸	مصرف مشروبات الکلی	Alcoholic (ALCO)	NO	بله	۲
			YES	خیر	
۱۹	دارای فرزند	Having Child (HCHI)	YES	بله	۲
			NO	خیر	
۲۰	دارای حیوانات خانگی	With Pet (PET)	NO	بلی	۲
			YES	خیر	
۲۱	سیاست	Policy (POL)	NO	بلی	۲

پایگاه داده به کار رفته در تحقیق فعلی شامل بیش از دو میلیون کاربر و به طور متوسط ۲۰ صفت صریح برای هر کاربر به اضافه‌ی دیتای ضمنی است.

کاربران می توانند فعال یا غیرفعال باشند. یک کاربر غیرفعال ممکن است شامل این‌ها باشد: کاربرانی که هیچ پروفایلی را ندیده اند، کاربرانی که با هیچ یک از کاربران دیگر ارتباط نداشته اند و کاربرانی که بیشتر از سه ماه است وارد حساب کاربری خود نشده اند. جدول(۲) نشان دهنده نحوه فیلتر کردن کاربران دیتاست مورد مطالعه تحقیق، برای کاهش تعداد آن‌ها است.

جدول۲: کاربران غیرفعال

تعداد کاربران	کاربرانی که در ۱۲ ماه گذشته وارد سیستم نشده اند		کاربرانی که با سایرین ارتباط نداشته اند	
	تعداد	درصد	تعداد	درصد
۲۰۹۲۵۸۵	۵۲۸۵۸۵	٪۲۵	۸۶۶۴۹۳	٪۴۱

در تحقیق فعلی پس از حذف کاربران غیرفعال، حجم دیتاست به حدود ۶۰۰هزار نفر خواهد رسید. از آنجاکه حجم دیتا در شبکه اجتماعی مورد مطالعه بالا می باشد، نمونه گیری از دیتاست انجام خواهد شد.با استفاده از فرمول کوکران[1] به محاسبه حجم نمونه آماری می پردازیم:

$$n = \frac{\frac{z^2 pq}{d^2}}{1 + 1/N(\frac{z^2 pq}{d^2} - 1)} \qquad (1-3)$$

که در آن n حجم نمونه ، N حجم جامعه آماری، z درصد خطای معیار ضریب اطمینان قابل قبول ، p نسبتی از جمعیت دارای صفت معین (جمعیت مردان)، q نسبتی از جمعیت فاقد صفت معین (جمعیت زنان)، d درجه اطمینان می باشد.با سطح خطای ۵درصد ،حجم نمونه برآورد شده برابر با ۳۸۲ کاربر می باشد.

در تحقیق فعلی با نرم افزار Weka ،صفاتی که هیچ تغییری در پیشنهادات ایجاد نمی‌کند، از لیست صفات خارج خواهیم کرد. دیتاهای دریافتی از شبکه اجتماعی مورد نظر برای تمامی صفات به فرم باینری(صفر و یک) می باشد. در ادامه جهت ملموس شدن صفات و مقادیر آنها، دیتا را از فرم باینری به اسمی(صنعت) و وصفی(تعداد فرزندان) تبدیل می کنیم.به محض اینکه دیتا ادغام و تبدیل شدند آنگاه به دو گروه تقسیم می‌گردند: مرد و زن. جدول(۳) تعداد کاربران زن و مرد را نشان داده است.

جدول۳: تعداد کاربران نمونه به تفکیک جنسیت

تعداد کاربران	کاربران مرد		تعداد کاربران زن	
	تعداد	درصد	تعداد	درصد
۳۸۲	۱۸۹	۴۹	۱۹۳	۵۱

در مرحله بعد، خوشه بندی کاربران را به تفکیک مرد و زن با الگوریتم خوشه بندی میانگین K انجام می دهیم. در مرحله بعد زمانی که کاربر جدیدی عضو سیستم می‌شود و در پروفایل خود صفاتش را تکمیل می‌کند، توسط الگوریتم SCClust، به کلاستری که بیشترین صفات مشترک را با وی دارد تخصیص می‌یابد.

[1] **Cochran**

یافته‌ها

جدول(۴) نشان می‌دهد که چه درصدی از صفات توسط کاربران در پروفایل هایشان ثبت می شود.

جدول۴:تعداد کل ثبت‌ها در هرصفت

صفات	درصد ثبت شدن‌های کل
گروه سنی	۱۴,۹۳
وضعیت تاهل	۱۰,۸۵
شخصیت	۹,۴۸
تحصیلات	۹,۳۲
قد	۶,۴۴
وزن	۵,۵۵
صنعت و پیشه	۵,۱۵
سطح مهارت شغلی	۴,۹۹
سیگار کشیدن	۴,۹۸
رنگ مو	۴,۴۴
رنگ چشم	۳,۷۹
دین	۳,۳۳
سطح اعتقادات	۳,۳۲
مصرف مشروبات الکلی	۳,۰۹
تمایل فرزند	۲,۸۳

دارای فرزند	۲,۴۲
ماه تولد	۲,۱۳
دارای حیوانات خانگی	۱,۴۸
رژیم غذایی	۱,۴۷
سیاست	۰,۰۱

صفاتی که در تصمیم گیری و انتخاب کاربران تاثیرگذار می باشد در جدول(۵) نشان داده شده است.

جدول۵: مرتبط ترین صفات کاربران

گروه سنی Age Category (AGE)	وضعیت تاهل Marital Status (STA)	شخصیت Character (CHA)	تحصیلات Education (EDU)	قد Length (LEN)	وزن Weight (WEI)	صنعت و پیشه Craft (CRA)
Teenager	Single	Introverted	Basic	Short	Thin	Unemployed
						Student
Young	Married		Elementary	Normal	Middle	Worker
Middle	Divorced	Extroverted				Employee
Old	Widow		Collegiate	Tall	Fat	Self
						Oder

همان طور که مشاهده می‌شود صفات انتخاب شده توسط نرم افزار Weka همان صفاتی هستند که در جدول(۵) بیشترین وزن را به خود اختصاص داده‌اند. پس با این تفاسیر از ۱۳ صفت دیگر کاربران چشم پوشی کرده و خوشه‌بندی کاربران را بر اساس این 7 صفت انجام می‌دهیم.

به منظور بررسی صحیح داده‌ها، صفات با مقادیر پیوسته(گروه سنی، قد و وزن)، گسسته سازی شده است. در جدول(۶) مقادیر گسسته سازی شده نشان داده شده است:

جدول ۶: گسسته سازی صفات

صفت	نماینده دسته	ابتدای بازه	انتهای بازه

گروه سنی	نوجوان	۱۶	۲۰
	جوان	۲۱	۳۵
	میانسال	۳۶	۵۰
	پیر	۵۱	۷۰
قد	کوتاه	۰	۱۶۰
	متوسط	۱۶۱	۱۸۰
	بلند	۱۸۱	۲۸۰
وزن	لاغر	۰	۵۰
	متوسط	۵۱	۹۰
	چاق	۹۱	۲۰۰

خوشه بندی با الگوریتم میانگین k

ورودی الگوریتم میانگین k، مشخص نمودن تعداد کلاسترها و همین طور داده‌ها و مقادیر صفاتشان می‌باشد. در تحقیق فعلی حجم کلاسترها را ۳ در نظر می‌گیریم. خروجی الگوریتم در شکل‌های(۲) و (۳) نشان داده شده است.

```
                    Cluster#
Attribute       Full Data         1          2          3
                  (189)        (95)       (56)       (38)
==========================================================================
AGE                          Teenager(42)  Middle(24)  Young(19)
STA                          Married(48)   Single(38)  Divorced(24)
CHA                          Introverted(75) Introverted(38) Extroverted(34)
EDU                          Elementary(43)  Collegiate(35)  Collegiate(20)
LEN                          Normal(50)    Normal(35)  Tall(27)
WEI                          Middle(64)    Thin(30)    Fat(19)
CRA                          Unemployed(45)  Employee(30)  Oder(15)

=== Model and evaluation on training set ===
```

Clustered Instances
1 50% (95)
2 30% (56)
3 20% (38)

شکل۲:خوشه بندی کاربران مرد با الگوریتم میانگین k

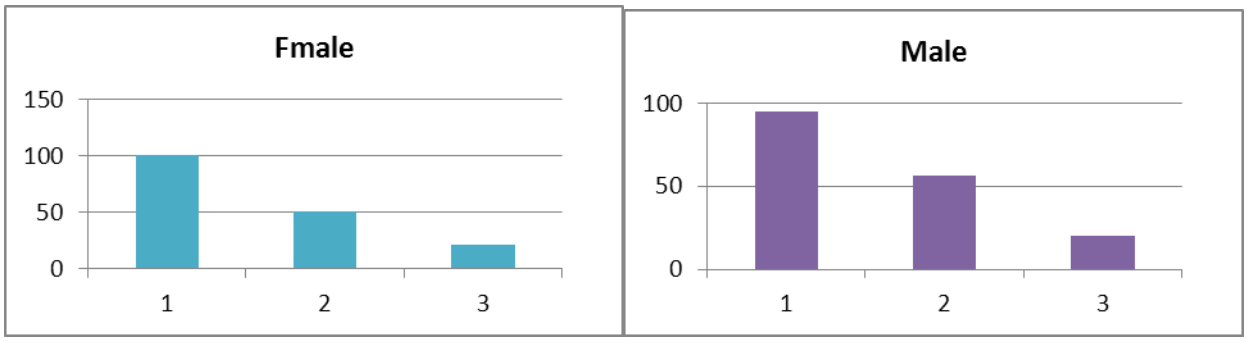
Attribute	Full Data (193)	Cluster# 1 (101)	2 (51)	3 (41)
AGE		Young(54)	Teenager(30)	Young(25)
STA		Single(54)	Married(32)	Single(31)
CHA		Introverted(85)	Extroverted(30)	Extroverted(40)
EDU		Elementary(53)	Collegiate(32)	Collegiate(24)
LEN		Normal(65)	Tall(30)	Normal(23)
WEI		Middle(56)	Thin(31)	Fat(30)
CRA		Oder(29)	Self(19)	Employee(20)

=== Model and evaluation on training set ===

Clustered Instances
1 53% (101)
2 26% (51)
3 21% (41)

شکل۳: خوشه بندی کاربران زن با الگوریتم میانگین k

تعداد کاربران هر کلاستر به تفکیک مرد و زن را در شکل(۴) می‌توان مشاهده کرد:



شکل۴: نمودار تعدادی کاربران هر کلاستر

معیارهای کیفیت خوشه بندی

۱.تشابه داخلی[۲]

جهت محاسبه تشابه داخلی برای کلاسترهای مرد و زن از فرمول زیر استفاده می کنیم:

$$InterSim = \frac{\sum_{i=1}^{c}\sum_{j=i+1}^{c} Clust_Sim_{i,j}}{0.5 \times c \times (c-1)} \qquad (1-4)$$

$$Clust_Sim_{i,j} = \begin{cases} 1 & , \quad IF \ a_{ki} = a_{kj} \\ 0 & , \quad Otherwise \end{cases} \qquad (2-4)$$

که در آن $Clust_Simi.j$ تشابه کلاستر i و j را نشان می دهد ، c بیانگر تعداد کلاسترها و a_k معرف مقدار صفت kام کلاسترهای i و j می باشد. گفتنی است که همیشه $0 < InterSim < 2m/(c-1)$ و m نمایشگر تعداد صفات می باشد. همانطور که اشاره شد، نزدیکی این معیار به صفر نشان دهنده خوشه بندی کارآمد می باشد و در تحقیق فعلی،این معیار برای ۳ کلاستر و ۷ صفت در محدوده ی صفر تا هفت قرار می گیرد و مقدار آن برای کلاسترهای مرد و زن به صورت زیر محاسبه خواهد شد:

$$InterSim\ (Cm) = \frac{\sum_{i=1}^{3}\sum_{j=i+1}^{3} Clust_Sim_{i,j}}{0.5 \times C \times (C-1)}$$
$$= \frac{Clust_Sim(1,2) + Clust_Sim(1,3) + Clust_Sim(2,3)}{0.5 \times 3 \times 2} = \frac{2+0+1}{0.5 \times 3 \times 2}$$
$$= 1$$

$$InterSim\ (Cf) = \frac{\sum_{i=1}^{3}\sum_{j=i+1}^{3} Clust_Sim_{i,j}}{0.5 \times C \times (C-1)}$$
$$= \frac{Clust_Sim(1,2) + Clust_Sim(1,3) + Clust_Sim(2,3)}{0.5 \times 3 \times 2} = \frac{0+3+2}{0.5 \times 3 \times 2}$$
$$= 1.67$$

۲.نرخ رضایت مندی محدودیت ها[۳]

[2] **Inter-Similarity**

در خوشه‌بندی، مقادیر صفات خوشه را مقادیر محدودیت‌دار گویند. نرخ رضایت محدودیت‌ها معیاری است که درصد کاربرانی که مقادیر صفاتشان با مقادیر صفات خوشه خود یکسان است را طبق فرمول زیر محاسبه می کند:

$$ConstSat(C) = \frac{1}{m \times c}\sum_{i=1}^{c}\left(\frac{\sum_{j=1}^{n} ConstSat(uj, ci)}{n}\right) \qquad (3-4)$$

$$ConstSat(uj, ci) = \begin{cases} 1 & , \ Constraint\ is\ satisfied \\ 0 & , \ Otherwise \end{cases} \qquad (4-4)$$

گفتنی است چنانچه نرخ رضایت‌مندی محدودیت، مقداری بالای ۵۰ درصد شود بیانگر این است که خوشه بندی دقیق و قابل قبولی اتفاق افتاده است. در تحقیق فعلی مقادیر زیر برای این معیار برآورد شده است:

$$ConstSat(Cm) = \frac{1}{7 \times 3}\left[\left(\frac{42+48+75+43+50+64+45}{95}\right) + \left(\frac{24+38+38+35+35+30+30}{56}\right) + \left(\frac{19+24+34+20+27+19+15}{38}\right)\right] = 0.18 + 0.20 + 0.20 = 0.58$$

$$ConstSat(Cf) = \frac{1}{7 \times 3}\left[\left(\frac{54+54+85+53+65+56+29}{101}\right) + \left(\frac{30+32+30+32+30+31+19}{51}\right) + \left(\frac{25+31+40+24+23+30+20}{41}\right)\right] = 0.19 + 0.19 + 0.22 = 0.60$$

همانطور که مشاهده می‌شود به طور متوسط ۵۸ درصد از کاربران مرد صفاتشان با صفات کلاستر تخصیص یافته‌ی خود همخوانی دارد و همین‌طور این عدد برای کاربران زن به ۶۰ درصد می‌رسد که نشان دهنده یک خوشه‌بندی خوب و قابل قبولی می‌باشد.

۳.تطبیق کاربران

برای تطبیق کاربران کلاستر مرد و زن در ابتدا تعداد صفات مشترک کلاسترهای مرد و زن را مشخص می‌کنیم. جدول(۷) تعداد صفات مشترک این کلاسترها را نمایش می‌دهد.

---

[3] **Constraints Satisfaction Rate**

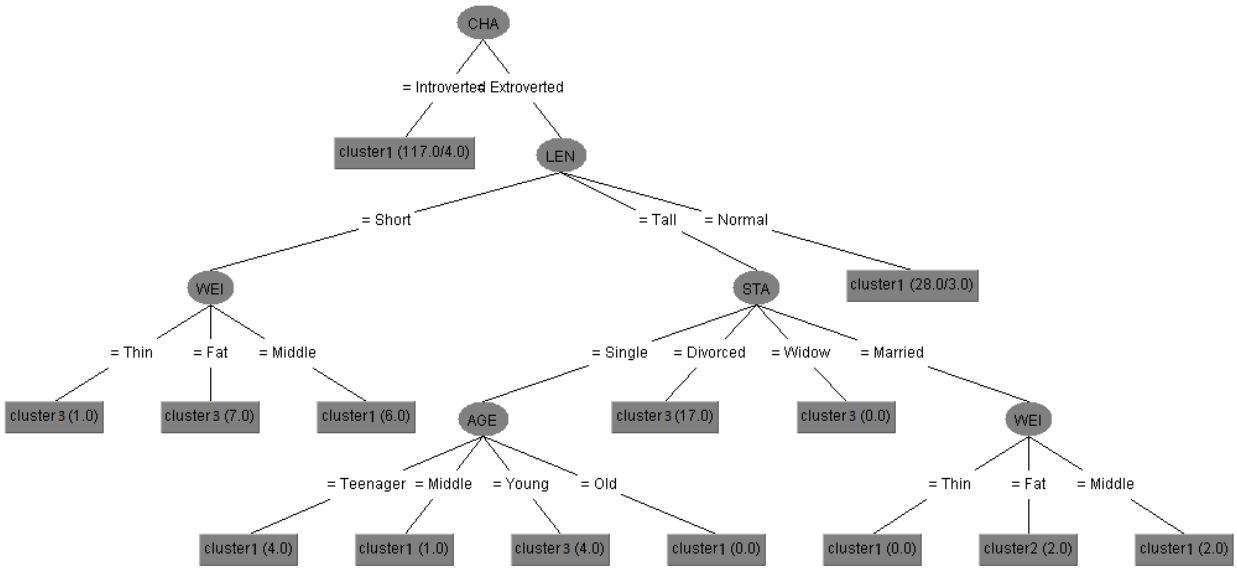
جدول۷:صفات مشترک کلاسترهای زن و مرد

تطبیق کلاسترها	تعداد صفات مشترک	نام صفات
1M1F	۴	Introverted,Elementary,Normal,Middle
1M2F	۲	Teenager,Married
1M3F	۱	Normal
2M1F	۵	Single,Introverted,Collegiate,Normal,Employe
2M2F	۲	Collegiate,Thin
2M3F	۳	Single,Collegiate,Normal
3M1F	۲	Young,Collegiate
3M2F	۳	Extroverted,Collegiate,Tall
3M3F	۴	Young,Extroverted,Collegiate,Fat

همانطور که مشخص است بهترین گزینه برای تطبیق کاربران بدین صورت می‌باشد که کاربران کلاستر ۱ مرد با کاربران کلاستر ۱ زن جفت شوند، کاربران کلاستر ۲ مرد با کاربران کلاستر ۱ زن جفت شوند و کاربران کلاستر ۳ مرد با کاربران کلاستر ۳ زن جفت شوند.

۴.تطبیق صفات کاربران مرد با کلاسترهای زن

برای این منظور صفات کاربران مرد را به عنوان متغیرهای وابسته و بهترین کلاستر زن پیشنهادی را به عنوان متغیر تصمیم گیری برای رسم درخت تصمیم در نرم افزار Weka وارد می کنیم. در ساخت درخت تصمیم از مجموعه داده‌های آموزشی مردان استفاده می شود و پس از پیاده سازی، صحت قابل قبولی برابر ۹۶ ٪ (با خطای کمتر از ۴ درصد) به دست آمد. خروجی این درخت در شکل (۵) نمایش داده شده است:



شکل۵:درخت تصمیم تطبیق صفات کاربران مرد با کلاسترهای زن

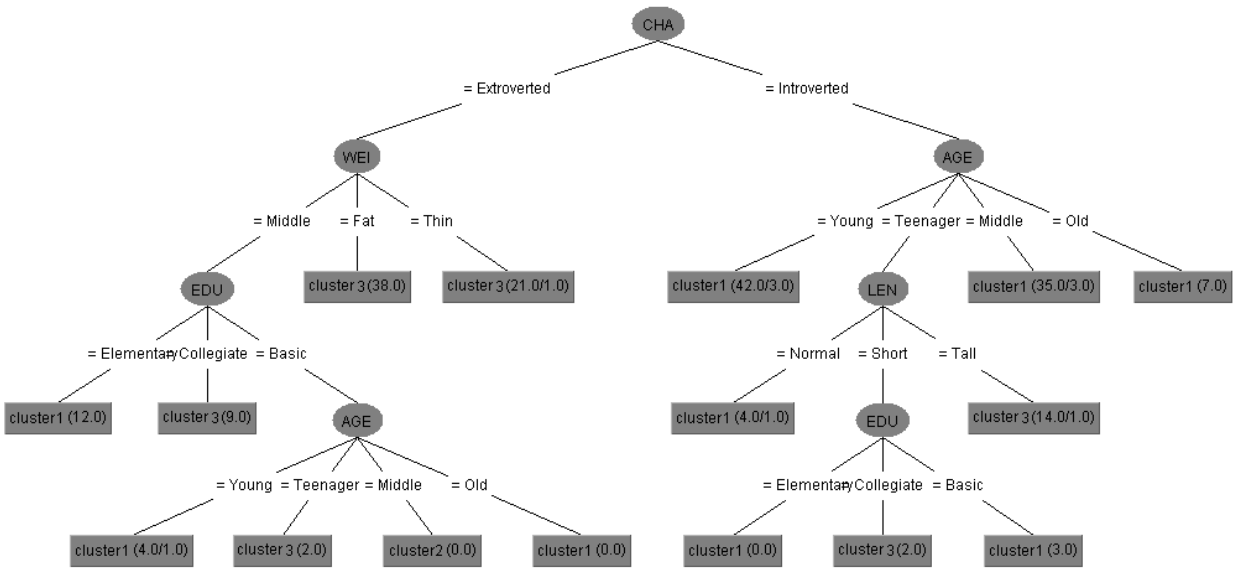
برای یک کاربر مرد برخی از پیش بینی‌ها تطبیق آینده با کلاسترهای زن در زیر بیان شده است:

- اگر کاربر مرد شخصیتی درون گرا داشته باشد آنگاه کلاستر ۱ کاربران زن به وی پیشنهاد خواهد شد.
- اگر کاربر مرد شخصیتی برون گرا داشته باشد و قد متوسط داشته باشد آنگاه کلاستر ۱ کاربران زن به وی پیشنهاد خواهد شد.
- اگر کاربر مرد شخصیتی برون گرا داشته باشد و کوتاه قد و لاغر باشد آنگاه کلاستر۳ کاربران زن به وی پیشنهاد خواهد شد.
- اگر کاربر مرد شخصیتی برون گرا داشته باشد و کوتاه قد و چاق باشد آنگاه کلاستر۳ کاربران زن به وی پیشنهاد خواهد شد.
- اگر کاربر مرد شخصیتی برون گرا داشته باشد و قد کوتاه و وزنی متوسط داشته باشد آنگاه کلاستر۱ کاربران زن به وی پیشنهاد خواهد شد.

- اگر کاربر مرد شخصیتی برون گرا داشته باشد و بلند قد و از همسر خود جدا شده باشد آنگاه کلاستر۳ کاربران زن به وی پیشنهاد خواهد شد . و دیگر حالت ها که در شکل مشخص است.

۵.تطبیق صفات کاربران زن با کلاسترهای مرد

در ادامه صفات کاربران زن را به عنوان متغیرهای وابسته و بهترین کلاستر مرد پیشنهادی را به عنوان متغیر تصمیم گیری برای رسم درخت تصمیم در نرم افزار Weka وارد می کنیم. در ساخت درخت تصمیم از مجموعه داده های آموزشی زنان استفاده می شود و پس از پیاده سازی، صحت قابل قبولی حدود ۹۵ ٪ (با خطای ۵ درصد) به دست آمد. خروجی این درخت در شکل(۶) نمایش داده شده است:



شکل۶:درخت تصمیم تطبیق صفات کاربران زن با کلاسترهای مرد

برای یک کاربر زن نیز برخی از پیش بینی ها تطبیق آینده با کلاسترهای مرد در زیر بیان شده است:

- اگر کاربر زن شخصیتی درون‌گرا داشته و جوان، میان سال یا پیر باشد آنگاه کلاستر ۱ کاربران مرد به وی پیشنهاد خواهد شد.

- اگر کاربر زن شخصیتی برون‌گرا داشته و لاغر یا چاق باشد آنگاه کلاستر ۳ کاربران مرد به وی پیشنهاد خواهد شد. دیگر حالت ها که در شکل مشخص است.

۳.تطبیق کاربر جدید

کاربر جدیدی که وارد سیستم می شود و جنسیت و مقادیر صفاتش را مشخص می کند، توسط الگوریتم SCClust صفاتش بررسی می شود و در یکی از کلاسترهای هم جنس خود که بیشترین تطابق صفات با وی را دارد ، قرار داده می شود.این الگوریتم در مرحله اول جنسیت کاربر را تشخیص داده و در ادامه کاربر را به کلاستری که بیشترین تعداد مقادیر مشترک با وی را دارد تخصیص می‌دهد. الگوریتم برای هر کلاستر تکرار می شود و کاربر از کلاستری به کلاستر دیگر جابجا شده تا نهایتا در بهترین کلاستر از کاربران هم جنس خود قرار می‌گیرد. شکل(۷) توصیفی از الگوریتم SCClust با جزئیات بیشتر است:

ورودی:

۱- داده های کاربر جدید U شامل مجموعه صفات $\{A_1, A_2, \dots, A_8\}$، و هر صفت مقداری به صورت $\{V_1, V_2, \dots, V_8\}$ می گیرد که در آن V_1 مرد یا زن بودن را نشان می دهد و V_2 تا V_8 هفت مقدار از صفات اصلی می باشند.

۲- کلاسترهای کاربران مرد $\{C_{1M}, C_{2M}, C_{3M}\}$ با مقادیر صفات $\{V_{1M}, V_{2M}, \dots, V_{8M}\}$.

۳- کلاسترهای کاربران زن $\{C_{1F}, C_{2F}, C_{3F}\}$ با مقادیر صفات $\{V_{1F}, V_{2F}, \dots, V_{8F}\}$.

خروجی:

تخصیص کاربر U به کلاستر C_C.

الگوریتم:

مقدار دهی اولیه:

۱- کاربر U را بخوان

۲- اگر $V_1=V_{1M}$ آنگاه، کاربر U را در C_{iM} که تعداد مقادیر مشترک را ماکزیمم می کند قرار بده .

۳- اگر $V_1 \neq V_{1M}$ آنگاه، کاربر U را در C_{iF} که تعداد مقادیر مشترک را ماکزیمم می کند قرار بده .

۴- <U.C_C> را بنویس،

تکرار:

۱- جابجایی= بلی(Move=True)،

۲- کاربرU را در C_j دیگری که تعداد مقادیر مشترک را ماکزیمم می کند قرار بده،

۳- اگر $C_j \neq C_C$ آنگاه

۴- <U.C_j> را بنویس،

۵- جابجایی= بلی(Move=True)،

۶- تا زمانیکه کاربر U دیگر جابجا نشود.

پایان

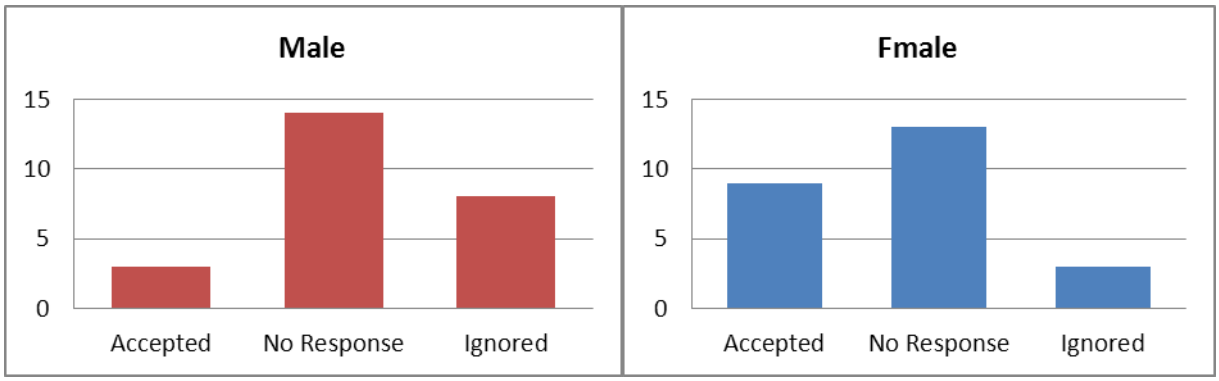
شکل۷: الگوریتم کلاسترینگ کاربر جدید

حال به کاربر جدید، با توجه به خوشه ای که در آن قرار گرفته، کاربران جنسیت مخالف، همانگونه که در پیش بینی درخت تصمیم اشاره شد، پیشنهاد می شود.

تجزیه و تحلیل داده ها

در این قسمت قصد داریم با بررسی داده های ضمنی کاربران، نمونه ای از ارتباطات موفق و ناموفق برقرار شده را مورد مطالعه قرار دهیم و نتایج آن را تحلیل و بررسی کنیم.

از آنجاکه این نوع از داده ها، مربوط به حریم خصوصی افراد می‌باشد لذا اطلاعات آن در دسترس محققین قرار نگرفته است. برای تحلیل رفتار کاربران، یک حساب کاربری مرد و یک حساب کاربری زن ایجاد گردید و از هر کدام ۲۵ درخواست دوستی برای کاربران مختلف سایت ارسال شد. شکل(۸) تعداد و درصد درخواست‌های پذیرفته شده، پاسخ داده نشده و رد شده برای هر نوع حساب کاربری را نشان می‌دهد.



شکل۸: تعداد درخواستهای پذیرفته شده، پاسخ داده نشده و رد شده

همان طور که مشاهده می شود، معمولا یک درخواست دوستی آمده از یک شخص ناشناس، قبول نمی‌شود. به ترتیب، تنها ۱۲ درصد درخواست دوستی حساب کاربری مرد از طرف زنان پذیرفته شده است یعنی از ۲۵ درخواست ارسالی تنها ۳ درخواست پذیرفته شده است و ۳۶ درصد درخواست دوستی حساب کاربری زن از طرف مردان پذیرفته شده است یعنی از ۲۵ درخواست ارسالی ۹ درخواست پذیرفته شده است.

در مرحله دوم صفات کاربرانی که درخواست دوستی را پذیرفته اند با صفات کاربر تطبیق می دهیم و تعداد صفات مشترکشان را بررسی می کنیم.جدول (۸) و (۹) تعداد صفات مشترک را نمایش می دهد:

تعداد کل صفات	تعداد صفات مشترک	روابط موفق
۷	۵	M1&F1
۷	۶	M1&F2
۷	۵	M1&F3

جدول۸: تعداد صفات مشترک کاربر مرد و زنان در حالت پذیرش

جدول۹:تعداد صفات مشترک کاربر زن و مردان در حالت پذیرش

تعداد کل صفات	تعداد صفات مشترک	روابط موفق
۷	۸	F1&M1
۷	۵	F1&M2
۷	۳	F1&M3
۷	۵	F1&M4
۷	۴	F1&M5
۷	۵	F1&M6
۷	۶	F1&M7
۷	۵	F1&M8
۷	۷	F1&M9

همان‌طور که از نتایج تست مشخص است، در حالت اول ۱۰۰٪ کاربرانی که با هم ارتباط موفق برقرار کرده اند حداقل ۵ صفت مشترک داشته اند. در حالت دوم نیز این عدد ۷۸٪ می باشد. نهایتاً صفات کاربرانی که درخواست دوستی را رد کرده اند با صفات کاربر تطبیق می‌دهیم و تعداد صفات مشترکشان را بررسی می کنیم.جدول (۱۰) و (۱۱) تعداد صفات مشترک را برای این دسته کاربران نمایش می دهد:

جدول۱۰: تعداد صفات مشترک کاربر مرد و زنان در حالت رد شده

روابط موفق	تعداد صفات مشترک	تعداد کل صفات
M1&F1	۱	۷
M1&F2	۰	۷
M1&F3	۱	۷
M1&F4	۱	۷
M1&F5	۲	۷
M1&F6	۳	۷
M1&F7	۱	۷
M1&F8	۱	۷

روابط موفق	تعداد صفات مشترک	تعداد کل صفات
F1&M1	۱	۷
F1&M2	۲	۷
F1&M3	۱	۷

جدول۱۱: تعداد صفات مشترک کاربر زن و مردان در حالت رد شده

در حالت اول ۸۸٪ کاربرانی که درخواست آن‌ها رد شده کمتر از ۲ صفت مشترک داشته اند. در حالت دوم نیز این عدد ۱۰۰٪ می باشد.

این نتایج گویای آن است که کلاسترینگ کاربران می تواند در تطبیق موفق، موثر باشد زیرا در کلاسترینگ کاربران ماکزیمم نمودن صفات مشترک بین دو کاربر شرط اصلی برای تطبیق بوده و کاربرانیکه صفات مشترک بیشتری با پارتنرهایشان دارند به هم پیشنهاد می شوند و از پیشنهاد دادن کاربرانیکه صفت مشترک کمی دارند جلوگیری می کند .

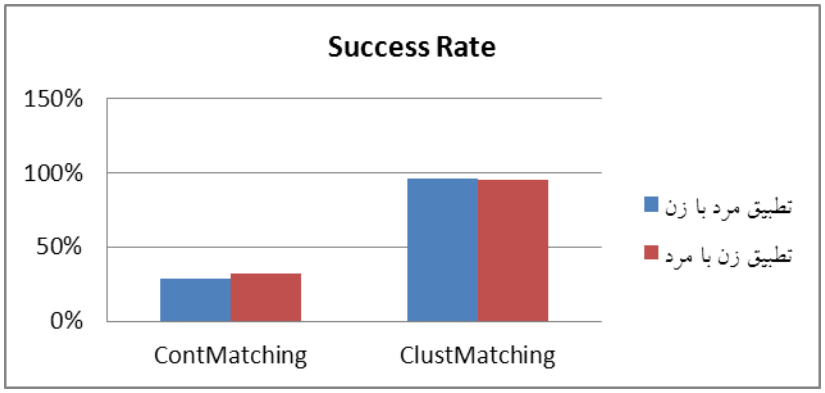
ارزیابی مدل پژوهش

در روش پیشنهادی توانستیم با استفاده از الگوریتم K-means کاربران را با هدف ماکزیمم کردن صفات مشترکشان خوشه بندی کنیم و نهایتا با استفاده از الگوریتم J48 ، کاربران را به بهترین کلاسترهای جنسیت مخالف تطبیق دهیم.

در ادامه، جهت تست مدل و محاسبه میزان تغییرات حاصل شده در دقت سیستم، با استفاده از دو تابع تعریف شده به نام های ContMatching و ClustMatching به محاسبه نرخ موفقیت[۴] تطبیق برای دو حالت :تطبیق دو به دو کاربران و تطبیق کلاستری کاربران، می پردازیم.قابل ذکر است که تابع ContMatching ،تمامی کاربران جنسیت مخالف را به کاربر پیشنهاد می کند و پس از تطبیق تمامی کاربران درصد موفقیت را در تطبیق محاسبه می کند در حالیکه تابع ClustMatching ،تنها بهترین کلاستر از جنسیت مخالف را به کاربر پیشنهاد می دهد و پس از تطبیق تمامی کاربران نرخ موفقیت را در تطبیق محاسبه می کند.در حالت اول، از نمونه مورد مطالعه، ۱۸۹ کاربر مرد را با ۱۹۳ کاربر زن، تطبیق دو به دو(ContMatching) می دهیم. در حالت دوم، از نمونه مورد مطالعه، ۱۸۹ کاربر مرد را به بهترین کلاسترهای زن پیشنهادی، تطبیق کلاستری(ClustMatching)

[4] **Success Rate**

می دهیم و نهایتا همین روال را برای کاربران زن نیز انجام می دهیم(تحلیل های دو تابع ذکر شده در پیوست قرار داده شده است). نتایج این تحلیل در شکل(۹) به تصویر کشیده شده است.



شکل۹:مقایسه نرخ موفقیت سیستم پیشنهادی با سیستم سنتی تطبیق

همانطور که نتایج نشان می دهد تطبیق کلاستری کاربران با افزایش چشم گیر نرخ موفقیت، از ۲۹ درصد به ۹۶ درصد(برای مردان) و از ۳۲ درصد به ۹۵ درصد(برای زنان)، باعث بهبود دقت سیستم، خواهد شد. اعداد گویای آن است که چنانچه از روش های کلاسترینگ و تطبیق اشاره شده در این تحقیق استفاده شود به رضایتمندی قابل قبولی نسبت به سیستم های سنتی دست خواهیم یافت.

نتیجه گیری

هدف اصلی در این تحقیق، بهبود و افزایش دقت و راندمان تطبیق و پیشنهاد دهی افراد در شبکه اجتماعی، در عین بهره وری از دیتاهای صریح کاربران با استفاده از تکنیک‌های خوشه بندی می باشد. این پژوهش یک سیستم تطبیق اجتماعی جدید بر اساس کلاستر بنا کرده است که باعث کاهش پیچیدگی فرآیند تطبیق شده و دقت را بالا می‌برد. این سیستم از دیتاهای صریح در مراحل مختلف در فرآیند برای اطمینان از پیشنهادات سازگار تر و غلبه بر برخی مشکلات موجود در سیستم‌های تطبیق اجتماعی امروزی استفاده می‌کند. این تحقیق به بررسی مسئله‌ی کلاستر کردن افراد در شبکه‌های اجتماعی پرداخته است و نهایتا با مقایسه نرخ

موفقیت سیستم پیشنهادی با سیستم های سنتی تطبیق کاربران، دقت و راندمان تطبیق افراد را به طرز قابل قبولی افزایش داده است.

منابع

E. Agichtein, Y. Liu, and J. Bian.Modeling information-seeker satisfactionin community question answering. ACM Transactions on KnowledgeDiscovery from Data (TKDD), 3(2):10, 2009.

S. Alsaleh, R. Nayak, Y. Xu, and L. Chen.Improving matching process insocial network using implicit and explicit user information.volume 6612,pages 313-320. Springer Computer Science, 2011. Proceedings of theAsia-Paci_c Web Conference (APWeb 2011) Lecture Notes in Computer Science.

S. Anand and B. Mobasher.Intelligent Techniques for Web Personaliza-tion.Springer, 20١5.

A. Kaplan and M. Haenlein. Users of the world, unite! the challengesand opportunities of social media. Business horizons, 53(1):59-68, 2010.

D. Knoke, S. Yang, and J. H. Kuklinski. Social network analysis, volume2. Sage Publications Los Angeles, CA, 2008.

#P. Krivitsky, M. Handcock, and A. Raftery.Representing degree distributiond, clastering, and homophily in social networks with latent cluster random effects models.Technical report no. 517 department of statistics university of washington , 11(1):1-18, 2007.

#P. Krishna Reddy, M. Kitsuregawa, P. Sreekanth, and S. Srinivasa Rao.A graph based approach to extract a neighborhood customer communityfor collaborative _ltering. Databases in Networked Information Systems,pages 188-200, 2002.

J. Kubica, A. Moore, D. Cohn, and J. Schneider.Finding underlyingconnections: A fast graph-based method for link analysis and collaborationqueries. In Machine learning international workshop, volume 20,page 392, 2003.

S. Kutty, L. Chen, and R. Nayak.A people to people recommendationsystem using tensor space models. In Proceedings of the 27th AnnualACM Symposium on Applied Computing, pages 187-192. ACM, 2012.

S. Kutty, R. Nayak, and Y. Li. Xml documents clustering using a tensorspace model.Advances in Knowledge Discovery and Data Mining, pages488-499, 2011.