Archive of SID

پیش‌بینی افراد خبره در شبکه‌های اجتماعی

روژین احمدی [۱]، سید امیر شیخ احمدی [۲]

[۱] کارشناسی ارشد مهندسی نرم‌افزار دانشگاه آزاد اسلامی واحد سنندج rozhinahmadi_sanandaj@yahoo.com

[۲] عضو هیات علمی گروه مهندسی کامپیوتر و فناوری اطلاعات دانشگاه آزاد اسلامی واحد سنندج، amir.sheikahmadi@gmail.com

چکیده:

امروزه تحلیل شبکه‌های اجتماعی به یکی از زمینه‌های تحقیقاتی جذاب تبدیل شده است. یکی از مسائل مهم در این زمینه تحقیقاتی، یافتن افراد خبره است. در این مسئله به نحوه یافتن یک فرد خبره در یک موضوع پرداخته می‌شود. در تحقیقات پیشین برای یافتن افراد خبره اغلب تنها به پروفایل آنها یا ارتباطات با دوستان و گروه‌ها یا ترکیبی از این موارد توجه شده است اما به این مسئله که چه فردی می‌تواند در آینده در دسته افراد خبره قرار گیرد توجه نشده است. بنابراین در این مقاله به پیش‌بینی افراد خبره پرداخته می‌شود. منظور از پیش‌بینی این است اینکه آیا فردی که جدیداً وارد سایت شده است مستعد خبره بودن هست یا نه. در روش پیشنهادی ابتدا به منظور برآورد نمره اولیه کارشناس برای هر فرد و انتخاب بالاترین رتبه‌بندی اشخاص به عنوان کاندیداها از اطلاعات شخصی افراد استفاده می‌کنیم؛ سپس برای حل مشکل شناسایی افراد خبره و جدا کردن آن‌ها از افراد، الگوریتم خوشه‌بندی امید ریاضی ارائه شده است که بر اساس فعالیت کاربران و بازخوردی که از کاربران دیگر دریافت می‌شود، افراد خبره از افراد دیگر جدا شده‌اند. در بخش بعدی برای استخراج الگوی افراد موفق، الگوریتم درخت تصمیم ID3 استفاده شده و در نهایت از ساختار پرسپترون برای پیش‌بینی احتمال تبدیل شدن کاربر جدید به فرد خبره استفاده شده است. شبیه‌سازی‌ها و آزمایشات انجام شده بر روی مجموعه داده کاربران، نشان می‌دهد که الگوریتم پیشنهادی از نظر دقت و زمان اجرا عملکرد بهتری نسبت به کارهای قبلی دارد.

کلمات کلیدی: شبکه‌های اجتماعی،پیش‌بینی افراد خبره،ساختار پرسپترون

۱. مقدمه

شبکه‌های اجتماعی معمولاً بر روی تعدادی از گره‌ها تعریف می‌شود که این گره باهم در ارتباط هستند. یکی از مهم‌ترین ویژگی‌های شبکه‌های اجتماعی به اشتراک گذاشتن دانش و اطلاعات است[۱]. به کمک این ویژگی افراد قادرند به صورت داوطلبانه مشاهدات خود از دنیای فیزیکی را منتشر کرده و اینکه علاوه بر ایجاد ارتباط، مشاهدات خود را با یکدیگر به طریقی کارا و سودمند به اشتراک گذارند. یافتن افراد خبره یکی از مهم‌ترین موضوعات برای استخراج داده‌ها از منبع اطلاعات شبکه‌های اجتماعی می‌باشد. یافتن افراد خبره هدفی برای شناسایی افراد با تخصص‌های مربوطه یا تجربه‌ای در مورد یک موضوع مورد نظر می‌باشد[۲].

مسئله شناسایی افراد خبره در تحقیقات متعددی مورد بررسی قرار گرفته است[۶-۲]، در اکثر این مطالعات از سه دیدگاه مسئله مورد بررسی قرار گرفته است. در دسته اول به پروفایل کاربران پرداخته شده است[۳]. دسته دوم ،روابط بین دوستان را به عنوان معیار انتخاب خبرگی در نظر گرفته است[۴]. در دسته آخر به گروه‌ها به عنوان معیار انتخاب توجه شده است[۵]. البته در تعدادی از کارها نیز به ترکیب این معیارها پرداخته شده است[۶].

در این مقاله، الگوریتم جدیدی برای پیش‌بینی افراد خبره در شبکه‌های اجتماعی با در نظر گرفتن هم‌زمان معیار پروفایل کاربران، روابط بین دوستان و گروه‌ها ارائه شده است. الگوریتم ارائه شده از دو مرحله اصلی تشکیل شده است. در مرحله اول ما به منظور برآورد نمره اولیه کارشناس برای هر فرد و انتخاب بالاترین رتبه‌بندی اشخاص به عنوان کاندیداها از اطلاعات شخصی افراد استفاده شده است؛ و سپس کاربری که جدیداً عضو شبکه اجتماعی مورد نظر شده است. آن‌ها را تست می‌کند و کارهای آن‌ها را در مدت زمان مشخصی دنبال می‌کند پست‌ها، علایق، توضیح‌ها و تائید دوستان در مورد فعالیت‌های که انجام می‌دهد. در کل در این مدت زمان چه فعالیت‌های انجام

www.SID.ir

Archive of SID

می‌دهند نسبت به کاربران دیگر مقایسه می‌شود. اگر فعالیت‌های زیاد و مفیدی داشته باشد نسبت به افراد دیگر سپس آن را با افراد خبره که در مرحله اول پیدا شده است مقایسه می‌شود که چقدر احتمال دارد این کاربر جز افراد خبره در شبکه اجتماعی باشد. اگر در حد آن‌های که جزء افراد خبره باشند، زودتر به لیست خبرگان این سایت اضافه می‌شود؛ که این یکی از کاربردها در این مسئله می‌باشد. ادامه مقاله به صورت زیر است:

در بخش ۲ به پیشینه تحقیق و کارهای انجام شده در رابطه با خبره‌یابی در شبکه‌های اجتماعی پرداخته خواهد شد در بخش ۳، به روش پیشنهادی روشی ارائه شده برای رتبه‌بندی کاربران، شناسایی فرد خبره پیش‌بینی فرد خبره در شبکه اجتماعی بررسی خواهد شد. در بخش۴ به بررسی نتایج بدست آمده و مقایسه آن‌ها پرداخته شده و در نهایت در بخش ۵ به نتیجه‌گیری و کارهای آتی اشاره خواهد شد.

۲. مروری بر کارهای انجام شده

مسئله خبره‌یابی در طی سالیان متمادی همواره چالش مورد بحث روز بوده است. روش‌های اولیه بیشتر بر اساس ورودی‌ها و داده‌های خود افراد در مورد مهارت‌ها و تخصص آن‌ها بوده است مارون[1] و همکاران [۷]. تخصص کاوی[2] بیشتر بر اساس توضیحات پروژه، پایگاه داده‌های منابع انسانی، مقالات حرفه-ای و تخصصی، کد برنامه یا فایل‌های خود افراد است. ایمیل نیز از این قاعده مستثنا نیست و به یکی از منابع متداول و مشهور برای تخصص کاوی بدل شده است زیرا اغلب برای ارتباطات در زمینه موضوعات کاری مورد بهره‌برداری قرار می‌گیرد بالوگ و دی رجکی[3] [۸]، با ظهور رسانه‌های اجتماعی، بسیاری از انواع کاربردهای آن معرفی شده است که به کاربران اجازه می‌دهد تا اطلاعات و دانش را از طریق این رسانه به اشتراک گذارند و باهم تعامل داشته باشند. این رسانه کاربران را قادر ساخته است تا در کامنت-گذاری، برچسب‌گذاری (تگ کردن)، پیوستن به گروه‌ها، با اشتراک‌گذاری، لایک‌کردن و پست‌گذاری، مشارکت کرده و به صورت فعال ایفای نقش کنند. تنوع انواع محتوا و نیز روابط کاربر با محتوا، پیشنهاد می‌کند که اطلاعات استخراج شده از این رسانه‌های اجتماعی می‌تواند حائز اهمیت باشد. برخلاف ایمیل یا فایل‌های محلی که شخصی بوده و برای مسائل حریم خصوصی بسیار حساس هستند، محتواها و روابط در این رسانه‌های اجتماعی عمومی است. داده‌ها از نظر ماهیت دینامیک بوده و تغییرات در فعالیت کاربر را با گذشت زمان، منعکس می‌کند. همچنین ساختار شبکه بر اساس روابط میان افراد، علاوه بر رابطه افراد با محتوا را منعکس می‌کند. به‌طورکلی سه روش اصلی برای یافتن خبره‌ها وجود دارد. روش اول بر آنالیز شبکه اجتماعی[4] (SNA) استوار است [۴]، و روش دوم بر اساس آنالیز محتوا است. [۹]، درنهایت، روش سوم ترکیبی از دو روش مذکور موسوم به روش هیبریدی[5] است[۱۰].

۲-۱ روش آنالیز شبکه اجتماعی (SNA): کمپل 33 و همکاران [۴]، به خبره یابی بر اساس ایمیل کاربران پرداختند .در این کار، از دو الگوریتم استفاده کردند: روش الگوریتم مبتنی بر محتوا که فقط متن ایمیل را در نظر می‌گیرد و الگوریتم دوم الگوریتم رتبه‌بندی مبتنی بر گراف (HLTS) که فقط متن و الگوهای ارتباطی را در نظر می‌گیرد .نتایج آن‌ها نشان داد که الگوریتم مبتنی بر-گراف، نتایج بهتری را در مقایسه با الگوریتم مبتنی بر محتوا ارائه کرد.

۲-۲ روش آنالیز محتوا: آسلای و همکاران [۹]، در مطالعه‌ای با عنوان "شبکه‌های مبتنی بر رقابت برای خبره‌یابی" به بررسی مسئله رتبه‌بندی کاربران در رابطه با تخصص آن‌ها پرداختند و ساختار شبکه جامعه خبره جدید با نام "شبکه‌های مبتنی بر رقابت برای خبره‌یابی" (CBEN) بر اساس اصل رقابت در میان پاسخ‌دهندگان به سؤالات پیشنهاد کردند. یافته‌های آنان نشان می‌دهد که توانایی برای شناسایی خبره‌ها به میزان بالایی به نوع فورام (forum) بستگی دارد که به‌طور مستقیم در ویژگی‌های ساختاری شبکه‌های تخصصی بازتاب یافته شده است. در این روش از سه مدل (۱) Asker-Replier networks (ARN) با یال‌های جهت‌دار از پرسش‌گر به پاسخ-دهنده سؤالات که بر اساس تعداد پاسخ‌ها ارزیابی می‌شود، (۲) Asker-Best Answerer Networks (ABAN) با یال‌های جهت‌دار از پرسش‌گر به بهترین پاسخ‌ها که بر اساس بهترین پاسخ‌ها ارزیابی می‌شود.(۳) ARN که در این مدل با تمامی پاسخ‌ها به‌طور یکسان برخورد شده و اطلاعات مربوط به " بهترین پاسخ‌ها" نادیده گرفته می‌شود. زمانی که پاسخی به‌عنوان " بهترین پاسخ" به یک سؤال خاص انتخاب می‌شود، با پاسخ‌های دیگر به همان سؤال مقایسه می‌شود. لذا، بین پاسخ‌های مربوط به سؤال مورد نظر نوعی رقابت ایجاد می‌شود. در همین راستا یک ساختار جدید به نام Competition-Based Expertise Network (CBEN) معرفی گردید که عملکرد بهتری را نسبت به مدل‌های ذکر شده با دقت بالا ارائه می‌کند.

[1] Maron
[2] Expertise Mining
[3] Balog & Derijke

[4] Social Network Analysis
[5] Hybrid Approach

www.SID.ir

Archive of SID

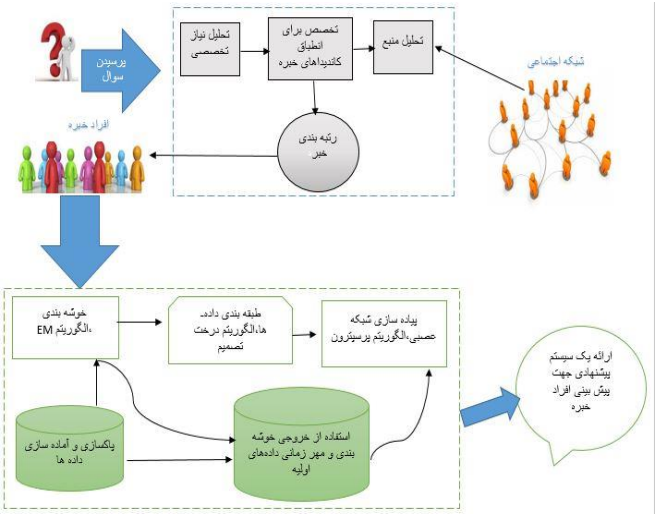
۲-۳ روش هیبرید: ریاحی ٦١ و همکاران [۱۰] در خبره‌یابی بر اساس وب سایت پرسش و پاسخ به بررسی سؤالات تازه پست شده پرداختند .در این مطالعه، از روش‌های مبتنی بر-کلمه ٦١ و مدل‌های موضوعی استفاده شده است .به طور کلی روش‌های مبتنی بر-کلمه از یک توزیع صاف برای تخمین احتمال یک تحقیق در یک مجموعه از پرونده‌ها استفاده می‌کند . مدل‌های موضوعی دارای سطح نمایشی اضافی هستند .پرونده‌ها در این مدل‌ها؛ مخلوطی از موضوعات هستند و موضوعات مخلوطی از کلمات هستند TF-IDF .یک ابزار اندازه‌گیری برای محاسبه اهمیت و ارتباط یک کلمه در پرونده بر اساس فراوانی کلمه در پرونده و عکس بخش پرونده‌ها حاوی کلمه در کل پرونده است .کلماتی که فقط در یک گروه کوچک از پرونده‌ها ظاهر می‌شوند، امتیازات tf-idf بالاتری نسبت به کلمات دیگر خواهند داشت -۱۱ TF-IDF .که در فرمول (۱-۲) آمده است، مجموعه پرونده‌Q ، کلمه w و پرونده:dϵQ

$$tfidf = fw,*\log\left(|Q|fw,Q\right) \qquad (2\text{-}1)$$

که $F_{W,d}$ تعداد دفعاتی است که w در d ظاهر می‌شود، Q اندازه توده پرونده-ها و$F_{W,d}$ مجموع تعداد پرونده‌هایی است که دارای کلمه w هستند

۳. روش پیشنهادی

مسئله رتبه‌بندی اعضای یک گروه اجتماعی را مطابق با سطح دانشی که آن‌ها در مورد یک موضوع خاصی دارند بررسی می‌شود. پس از چنین رتبه‌بندی،خبره‌ها با امتیاز برتری انتخاب می‌شوند. این خبره‌ها می‌توانند بسیاری از نیازهای مختلف را توسعه دهند، به سؤالات واقعی (جستجوهای مربوط به جستجوی-جمعیت) پاسخ می‌دهند، سپس برای جداسازی افراد خبره از بقیه افراد بنا به فعالیت‌های انجام شده در شبکه اجتماعی از الگوریتم خوشه‌بندی استفاده شده است، برای مشخصه سازی راهی که افراد برتر رفته‌اند از الگوریتم درخت تصمیم استفاده شده تا برای ارائه سیستم معلوم شود چه کسی شایستگی خبره بودن هست و چه کسی نیست. در بخش‌های هر یک از اجزای آن به تفصیل تشریح می‌گردد.



شکل۳-۱ شمای کلی از پیش‌بینی افراد خبره

۳-۱ تحلیل منابع

با توجه به نقش مهم و فزاینده شبکه‌های اجتماعی، این شبکه‌ها در انتقال اطلاعات از راندمان و بازدهی بالایی برخوردار هستند. شبکه‌های اجتماعی ابزاری برای ارتباطات به‌موقع و تعاملی هستند که مبادله دیالوگ و محتوا را در میان مصرف‌کنندگان و سازندگان پیام‌ها سهولت می‌بخشند [۱۱]. شبکه‌های اجتماعی فرصت‌هایی را برای گفتمان و تعامل دو-سویه در میان افراد و سازمان‌ها برقرار می‌کنند [۱۲] اطلاعات در قالب پست‌ها، کامنت‌ها، ایده و آرا بر روی شبکه‌های اجتماعی نمایان می‌شود. همچنین لینک‌هایی که به پروفایل‌ها، صفحات وب و دیگر پایگاه‌های داده‌های کاربر متصل می‌شود، نیز در مجموع منابع کاربر را تشکیل می‌دهند.

۳-۲ نیاز به تخصص و پردازش منبع

به منظور انجام فرایند انتخاب فرد خبره، دو مورد، یعنی منابع اجتماعی مربوط به کاندیداهای خبره و نیاز به تخصص، مورد تجزیه و تحلیل قرار گیرد. آنالیز بر روی نیازها و منابع به صورت متقارن انجام می‌گیرد. گام اول استخراج داده‌های اجتماعی از پلاتفرم‌های مختلف از طریق API های آن‌هاست. بدین منظور، داده‌های مربوطه با استفاده از پلاتفرم crowd-searcher برای جمع‌آوری نشانه‌های احراز هویت کاربران و اجزاءهای دسترسی به حریم خصوصی استفاده گردید است. سپس برای منابع مورد نظر، ماژول Resource Extraction عمل آنالیز را انجام داده است. با استفاده از اصطلاحات کاربردی پلاتفرم‌های اجتماعی مورد، هیچ اطلاعاتی بر روی سیستم‌های ذخیره ثانویه، ذخیره نمی‌شود. محتوای منابع در حالت

www.SID.ir

Archive of SID

عادی از متن تشکیل شده است، اما اغلب شامل URL های لینک مقالات، وبلاگ‌ها، پست‌ها کامنت‌ها و صفحات وب است.

۳-۳ مطابقت نیاز به تخصص با خبره‌های کاندیدا

این روش بر اساس یک مدل فضای برداری است که در آن منابع، هویت‌های مربوطه و نیازهای تخصصی، در همان فضا نشان داده شده است. با توجه به نیاز تخصصی و یک مجموعه از منابع مجموعه منابع مربوطه با محاسبه ارتباط هر منبع تولید می‌شود[3].

رتبه‌بندی خبره، پاک‌سازی و آماده‌سازی داده‌ها بر روی پایگاه داده صورت می‌گیرد در ابتدا جدول پست با جدول سابقه پست ادغام شده است. این روند در مورد تگ‌ها نیز صدق می‌کند و به جدول پست منتقل می‌گردد. طبق جدول کاربر، تعداد پست‌ها، کامنت‌ها و مدال‌های هر کاربر جمع و در رکوردی دیگر در جدول کاربر اضافه می‌گردد. با استفاده از تکنیک پیش‌بینی، تعدادی از خصیصه‌هایی که تأثیری در تعیین خبره بودن فرد ندارند و یا رکوردهایی که دارای نویز هستند، حذف می‌گردد. برای حل مشکل بزرگ بودن فضای مسئله از گسسته سازی داده‌های استفاده می‌شود. در این مرحله داده‌ها برای داده‌کاوی آماده‌سازی شده است.

۳-۴ خوشه‌بندی در مدل پیشنهادی

بر اساس بررسی مقالات و تحقیقات صورت گرفته درزمینه خبره یابی، دو انگیزه اصلی برای یافتن افراد خبره وجود دارد.

- به عنوان منبع اطلاعات
- به عنوان کسی که می‌تواند به فعالیت سازمانی یا اجتماعی را انجام دهد.

این طبقه‌بندی هر چند که فازی بوده و گاهی اوقات همپوشانی دارند، اما ثابت شده است که در آنالیز اهداف خبره‌یاب‌های خودکار و یا سیستم‌های خبره‌یاب، بسیار سودمند و مفید است. بسیاری از افراد ممکن است که در جستجوی افراد خبره به عنوان منبع اطلاعاتی برای تکمیل و یا جایگزین کردن منابع دیگر از جمله پرونده‌ها و پایگاه‌های داده در سناریوهای مختلف باشند.

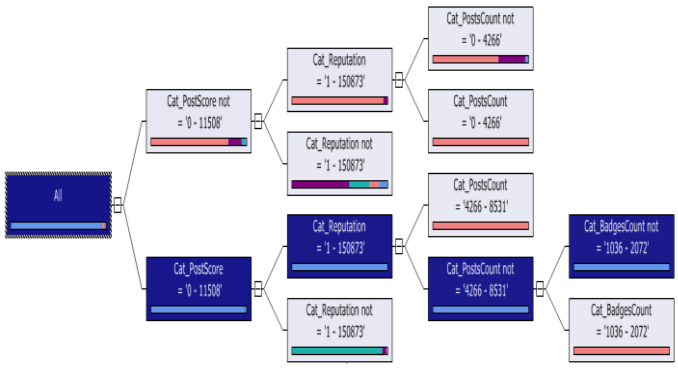
در خوشه‌بندی داده‌های بدست آمده از شبکه stackoverflow، از روش خوشه‌بندی مبتنی- بر دانسته (تراکم نقاط) استفاده شده است. در این روش اساس پارتیشن‌بندی مبتنی بر فاصله بین آن‌ها و اثر همسایگی بین نقاط است. در این مطالعه ۶۸۸۹ کاربر به‌عنوان مجموعه داده اولیه برای خوشه‌بندی ثبت گردیده است. خوشه‌بندی بر اساس اهم فعالیت‌های کاربران، یعنی لایک‌های مربوط به پست‌های آنان، امتیازات پست‌ها، تعداد رای‌های اختصاص یافته مربوط به جذابیت پست‌ها، لایک‌های کامنت‌ها، امتیازات پاسخ‌ها به سؤالات، مدال‌ها انجام‌شده است؛ که بر این اساس، مجموعه داده‌ها به چهار دسته «خبره قوی، خبره متوسط، خبره ضعیف، نا خبره» خوشه‌بندی می‌گردد جدول (۳-۱) خوشه۱ که کاربران عادی هستند تعداد آن‌ها از همه بیشتر است زیرا کاربران زیادی هستند که فقط عضو سایت می‌شوند و دیگر فعالیتی انجام نمی‌دهند، تعداد کاربرانی که در خوشه۴ هستند که جز کاربران خبره هستند از همه کمتر می‌باشد.

جدول ۳-۱. جدول خوشه‌بندی داده‌ها

خوشه	تعداد کاربران
خوشه ۱	۳۷۷۷
خوشه ۲	۱۴۲۶
خوشه ۳	۱۰۲۲
خوشه ۴	۶۶۲

۳-۵ طبقه‌بندی با استفاده از درخت تصمیم

از درخت تصمیم برای استخراج الگوی افراد موفق استفاده می‌گردد. نتایج بدست آمده به صورت (۲-۲) نشان داده شده است. از مراحل خوشه-بندی برای ساخت درخت تصمیم استفاده شده است. بدین منظور از الگوریتم id3 استفاده می‌شود..



شکل ۳-۲ درخت تصمیم با استفاده از الگوریتم ID3 شبکه‌های عصبی در مدل پیشنهادی

www.SID.ir

Archive of SID

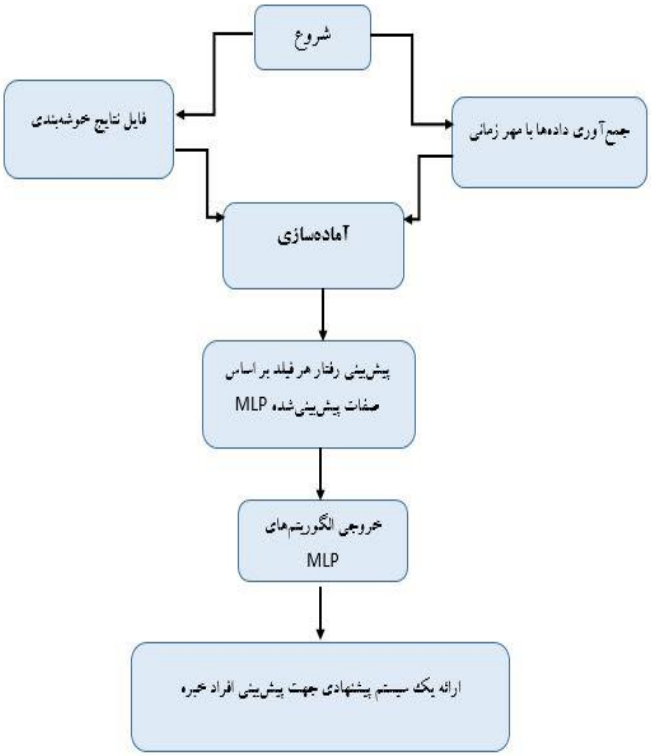
ابتدا افراد موفق یا خبره جداسازی شده، سپس الگوی موفقیت آن‌ها اکتشاف گردید. سپس بررسی می‌گردد که در خوشه ۳ چه کاربرانی قابلیت لازم برای عضویت در خوشه ۴ را دارا هستند. بدین منظور، از شبکه عصبی استفاده می‌گردد. در این روش، ساختار MLP مورد استفاده قرار گرفته است. روش پژوهش بدین منوال است که کاربران، در چه زمانی فعالیت‌های خود را انجام داده‌اند. برای انجام این کار از خوشه ۲ و ۴ برای آموزش شبکه استفاده گردید که هم افراد ضعیف و هم افراد خبره را مدنظر قرار می‌دهد. چون در این بخش از شبیه‌سازی، بازه زمانی از اهمیت به سزایی برخوردار است، لذا از مهر زمانی استفاده می‌گردد؛ که به دین خاطر، فعالیت‌های دو ماه اخیر کاربران مدنظر قرار می‌گیرد

۳-۶ شبکه عصبی در مدل پیشنهادی

در چنین پژوهش‌هایی ما نیاز به پیش‌بینی‌هایی با دقت و صحت بالا داریم .شبکه‌ی عصبی با بهره‌گیری از ساختار عصبی انسان قادر خواهد بود شبکه‌ای را آموزش داده و جهت پیش‌بینی از آن بهره گیرد .دقت این شبکه بسیار بالاست چراکه شبکه‌ی عصبی به شدت به خطا حساس است و می‌تواند هر خطایی را در مراحل اولیه‌ی شکل گیری، شناسایی نموده و حذف کند

۳-۶-۱پردازش و گزارش اولیه

در این بخش، داده‌های به دست آمده از خوشه‌بندی مورد پردازش قرار می‌گیرد. بدین منظور که معلوم گردد که هر فرد جزء کدام خوشه قرار می‌گیرد. سپس از داده‌های اولیه، دوباره استفاده می‌گردد. لازم به ذکر است که در این مرحله مهرزمانی داده‌ها نیز برای هر کاربر در نظر گرفته می‌شود تا معلوم گردد که فعالیت هر کاربر در زمان خاص به چه میزان بوده است. پس از مرحله شبیه‌سازی، بدین‌صورت فرض می‌گردد که درصورتی‌که شبکه عصبی جواب یک را در جلوی مشخصات اختصاص دهد، فرد به‌عنوان خبره در نظر گرفته می‌شود و درصورتی‌که عدد دو را اختصاص دهد، فرد کاربر در دسته افراد ضعیف دسته‌بندی می‌گردد. انتخاب مهر زمانی دو ماه بدین سبب است که شبکه اجتماعی مورد بررسی (stackoverflow) ازجمله شبکه‌های تخصصی پرسش و پاسخ در زمینه‌های مختلف بوده و لذا میزان ارجاع فرد به این شبکه‌ها عموماً پایین است و یا آنکه در طی مدت زمان کوتاه عموماً افراد کاربر، فعالیت چشمگیری از خود نشان نمی‌دهند. به همین خاطر مهر زمانی دو ماهه برای تمامی کاربران در نظر گرفته شد تا اثر این عامل کاهش یافته و تعداد افراد بیشتری در سیکل محاسبات قرار گیرد. در شکل (۳-۶) ترتیب مراحل شبیه‌سازی را نشان می‌دهد.



شکل۳-۳. ترتیب مراحل شبیه‌سازی عصبی

۳-۶-۲ ترتیب مراحل شبیه‌سازی عصبی

در شبکه عصبی از ساختارMLP استفاده می‌شود. در این مطالعه هر دو نوع ساختار کار یکسانی را انجام می‌دهند، اما به منظور افزایش دقت در نوع روش پیشنهادی، یک‌بار با استفاده از ساختار mlp شبیه سازی انجام گرفت. به گونه‌ای که اساس کارکرد آن مبتنی بر زمان باشد تا مقایسه‌ها بر اساس مبنای زمانی صورت گیرد. برای آزمایش و پیش بینی افراد در خوشه ۳ استفاده شده است. این خوشه دسته افراد متوسط را تشکیل می‌دهد. جهت پیش‌بینی، افراد خبره، بین خوشه ۳ و خوشه ۴ و بین خوشه ۳ و خوشه ۲ مقایسه صورت می‌گیرد. در صورتی که نتایج شبیه سازی خوشه ۳ را جز خوشه ۴ در نظر می‌گیرد، پیش بینی می‌شود که خوشه ۳ لیست افراد خبره را در آینده تشکیل می‌دهد، اما در صورتی که خوشه ۳ جز خوشه ۴ محسوب گردد، جز افراد ناخبره در نظر گرفته می‌شود از لیست پیش بینی خارج می‌گردد

۳-۶-۳ پیاده‌سازی MLP

در پیاده‌سازی با شبکه‌ی عصبیmlp ، طی سه مرحله و با تعداد لایه پنهان‌های مختلف مورد آزمایش قرار، گرفته است که در این پروژه به این

www.SID.ir

Archive of SID

ترتیب تعداد لایه‌ها را بررسی کرده‌ایم. لذا می‌توان با توجه به نتایج حاصل، بهترین انتخاب را برای تنظیمات آزمایش تعیین کرد. دستور لایه‌بندی و تعداد به صورت شکل (۳-۴) می‌باشد

Toil Black beads

```
net=newff(intrain', outtrain', [6 5 1]);
net=train(net, intrain', outtrain');
y = round(sim(net, intest'));
y = y';
export=zeros(56,7);
for i=1:56
export(i,1:6)=intest(i,1:6);
export(i,7)=y(i);
end
```

شکل ۳-۴. شبه کد برای لایه‌بندی mlp

۴. نتایج بدست آمده

آزمودن شبکه‌ی به دست آمده با استفاده از داده‌های اولیه انجام می-گیرد، چراکه ما از قبل داده‌های مربوط به این دوره را داشتیم و کافی است این داده‌ها را با نتایج حاصل از شبکه‌ی عصبی مقایسه کنیم، چنانچه اختلاف نتایج در حد قابل قبولی باشد، می‌توان نتایج را برای دوره‌های آتی با فرض عدم وجود تغییرات عمده پیش‌بینی کرد. جهت آموزش شبکه‌ی عصبی از جعبه ابزار شبکه عصبی همراه با کدنویسی‌های شبکه استفاده کردیم

۴-۱ مجموعه داده‌ها

داده‌های مورد استفاده در الگوریتم‌های پیشنهادی از سایت stackoverflow گرفته‌شده است. این اطلاعات مربوط به اعضای کاربر سایت بوده که شامل ۶ جدول اطلاعات کاربر[۶]، پست‌ها، کامنت‌ها[۷]، سابقه پست‌ها[۸]، مدال‌ها[۹]، تگ‌ها[۱۰] (User,Post,tag) از فعالیت‌های آن‌ها است. جداول بدست آمده از سایت، مورد مطالعه و بررسی قرار گرفت که تعداد ثبت شده از طریق سایت، برابر با ۶۸۸۹ رکورد بدست آمد و جداول دیگر، فعالیت همان ۶۸۸۹ کاربر ثبت شده است. لازم به ذکر است که جهت حفظ حریم خصوصی افراد، اطلاعات هویتی، تماس، آدرس و سایر اطلاعاتی که به نحوی بتواند به شناسایی افراد منجر شود از حوزه این پژوهش خارج بوده و لذا مستندات مربوطه ارائه نشده است. جدول (۳-۱) لیست جداول با تعداد رکوردها می‌باشد.

جدول ۴-۱. مجموعه داده‌های بدست آمده از شبکه stackoverflow

منبع	رکورد
User	۶۸۸۹
Comment	۱۲۵۰۶۵۱
Post	۲۹۹۳۹۳۵
Post History	۳۴۸۵
Tag	۲۳۰۷
Badges	۷۷۹۳۱۴

۴-۲ تنظیم پارامترها

در روش ارائه شده پارامترهای زیادی برای تنظیم وجود ندارد. تنها پارامترهای قابل تنظیم در این روش پارامترهای خوشه‌بندی و تعداد نرون‌ها می‌باشد؛ که مشخص کننده تعداد افراد خبره و افرادی که در آینده جز افراد خبره قرار می‌گیرند، که در این روش برای هر دوی آنها تعداد ۶۸۸۹ کاربر انتخاب شده است که این تعداد از طریق آزمون و خطا بدست آمده است در تمامی نتایج که در کارهای انجام شده صرفاً جهت شناسایی افراد خبره تنظیم شده است نه پیش‌بینی افراد خبره. برای بررسی پیش‌بینی کردن کاربران از ساختار پرسپترون استفاده شده است . تمام شبیه‌سازی‌های انجام شده در این پژوهش بر روی یک سیستم خانگی با مشخصات سی پی یو intel core i7 با فرکانس ۱،۷ گیگاهرتز برای هر هسته، حافظه اصلی ۴ گیگابایت انجام گرفته است.

۴-۳ نتایج شبیه‌سازی

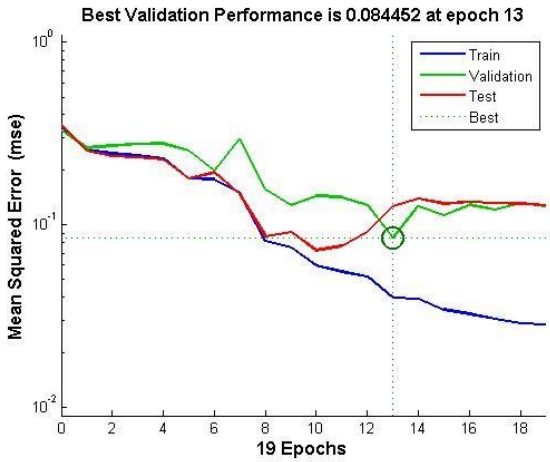
اولین مورد در رابطه با نتایج حاصل از شبکه‌ی عصبی به عملکرد شبکه حین آموزش مربوط می‌شود. شکل (۴-۱) نشان‌دهنده‌ی عملکرد شبکه‌ی عصبی حاصله حین فرآیند آموزش دیدن با استفاده از داده‌های اولیه است. در این شکل چهار نمودار داریم. محور افقی دوره‌های آموزش شبکه را نشان می‌دهد، تعداد دوره‌های آموزش ۱۹ دوره است که البته در هر بار اجرا این

[6] User
[7] Comment
[8] Posthistory
[9] Badges
[10] Tag

www.SID.ir

Archive of SID

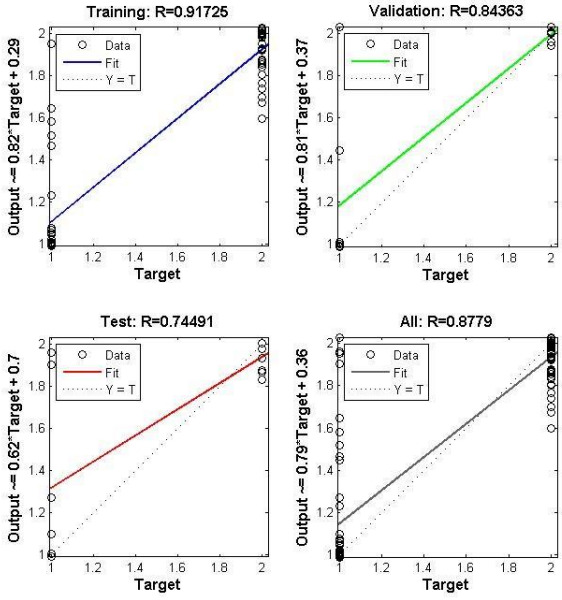
مقدار متفاوت است، این ناشی از ماهیت تصادفی و احتمالاتی شبکه‌های عصبی است. محور عمودی نیز مجذور توان دوم خطاهاست، این نوع سنجش خطا کمک می‌کند تا اختلاف از مقدار مورد انتظار را صرف‌نظر از مثبت یا منفی بودن آن ببینیم. از آنجا که تعداد دوره‌های آموزش شبکه نسبتاً مناسب است، سعی کردیم خطا نیز به مقداری در حد 10^{-2} کاهش یابد. این خط را با نام خط هدف می‌شناسیم. قاعدتاً این خط ایده‌آل‌ترین حالت آموزش شبکه است، اما دستیابی به آن در چنین شبکه‌ای میسر نشد. این از اعتبار کار نمی‌کاهد زیرا مشاهده می‌شود که اختلاف داده‌های حاصل از شبکه‌ی آموزشی با بیشتر شدن تعداد دوره‌های آموزش (خط آبی رنگ) به مقداری در حد ۰,۰۱ می‌رسد که این در نوع خود بسیار قابل توجه بوده و شبکه‌ای با دقت بالا را نوید می‌دهد. داده‌های آزمون (خط قرمز رنگ) و اعتبار سنجی (خط سبز رنگ) نیز دقت مناسبی دارند. البته این خط ارتباط چندانی با هدف نهایی ما از آموزش شبکه ندارند و ترسیم آن‌ها در عملکرد شبکه صرفاً جهت نمایش خط سیر ثابت آن‌ها بوده که این موضوع خود نشانگر عدم وجود اطلاعات پرت در شبکه است. مشاهده می‌شود در دوره‌ی ۱۳ بهترین نتیجه حاصل شده است، یعنی شبکه‌ی آموزشی در بالاترین حد دقت و صحت خود قرار دارند. ممکن است این سؤال مطرح شود که چرا آموزش شبکه تا دوره‌های بسیار بالاتری ادامه نیافته است، پاسخ این سؤال بسیار ساده است، با افزایش دوره‌های آموزش، شبکه به دقت بالا دست یافته است و چون در دوره‌ی ۱۳ و پیش از ۱۹ به بهترین نتایج رسیده، دیگر نیازی به طولانی‌تر شدن دوره‌های آموزش نیست و حتی با افزایش این دوره‌ها به نتایجی بهتر از این نتیجه نخواهیم رسید. مورد بعدی ماهیت تصادفی آموزش شبکه است بدین صورت که ممکن است در اجرای دیگری از همین ساختار نتایج اندکی تفاوت داشته باشد، اما روند کلی همچنان پابرجاست.



شکل ۴-۱. عملکرد شبکه‌ی عصبی حاصل حین فرآیند آموزش

آخرین مورد به رگرسیون اطلاعات حاصل از آموزش شبکه اختصاص دارد. در شکل ۴-۳ چهار نمودار مشاهده می‌شود، اولین آن‌ها (بالا سمت چپ) به داده‌های آموزشی حاصل از شبکه‌ی آموزش دیده مربوط می‌شود. مشاهده می‌شود که رگرسیون این داده‌ها به مقدار یک بسیار نزدیک است (۰,۹۱۷۲۵) که این گویای دقت بالای آموزش شبکه است. در این‌گونه موارد خروجی تحلیل رگرسیون را که از مقایسه‌ی داده‌های سال واقعی و حاصل از شبکه به دست می‌آید با خط مبنا مقایسه می‌کنند. شیب این خط مبنا برابر با مقدار عددی یک است و هرچه داده‌ها به این خط نزدیک‌تر باشند رگرسیون حاصل نیز پراکندگی بهتری در نزدیکی بهترین مقدار ممکن دارند. گرچه داده‌های اعتبارسنجی و آزمون رگرسیون کمتری دارند اما باز هم پراکندگی مناسب آن‌ها در مبدأ بیانگر استحکام شبکه است، می‌توان این مورد را در نمودار چهارم (پایین سمت راست) مشاهده کرد. این نمودار نشان می‌دهد که در کل آموزش شبکه مناسب بوده و رگرسیونی در حد ۰,۸۸ دارد.

حال که نشان دادیم شبکه به خوبی آموزش داده شده است می‌توانیم از آن جهت انجام هرگونه تحلیلی استفاده کنیم.



شکل ۴-۲. رگرسیون شبکه‌ی آموزشی برای داده‌های آموزشی، اعتبارسنجی، آزمون و کلی.

در مسائل مدل‌سازی و بهینه‌سازی تابعی به نام performance function معرفی می‌شود و هدف نهایی به حداقل رسانیدن مقدار این تابع است.

www.SID.ir

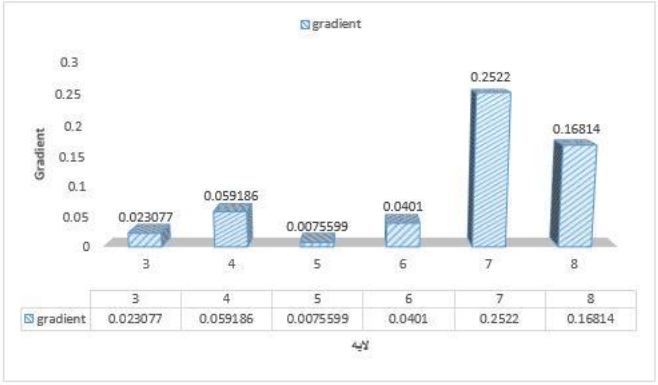
Archive of SID

$$performance = 1/(p*q)\sum_{i=1}^{p}\sum_{j=1}^{q}(output(i,j) - target(i,j)) \quad (۴-۱)$$

در رابطه فوق output خروجی‌ای است که شبکه‌ی عصبی پیش‌بینی می‌کند و target مقدار واقعی خروجی سیستم است. هدف شبکه‌ی عصبی این است که به نقطه مینیمم تابع فوق دست پیدا کند؛ بنابراین هرچقدر مقدار performance کمتر باشد نتایج بهتری خواهیم داشت. در نقطه‌ی مینیمم تابع فوق گرادیان صفر است، بنابراین یک روش دیگر تحلیل مسئله این است که شبکه‌ی عصبی به نقطه‌ای دست پیدا یابد که گرادیان صفر (کمترین مقدار) را داشته باشد.

۴-۴ مقایسه مقدار لایه‌ها با یکدیگر

مهم‌ترین مشخصه در تحلیل شبکه‌های عصبی، عملکرد یا performance است، پس از آن نیز گرادیان اهمیت فراوانی دارد. در نمودار ۴-۱ رابطه‌ی میان گرادیان و استفاده از لایه‌های مختلف به عنوان لایه‌ی پنهان شبکه‌ی عصبی نشان داده شده است. این رابطه تا اندازه‌ی زیادی تجربی است، یعنی در تلاشیم با تغیر تعداد نوع لایه و تعداد نرون‌های هر لایه به بهترین مشخصه برای شبکه‌ی عصبی دست بیابیم. مشاهده می‌شود کمترین مقدار گرادیان مربوطه به زمان استفاده از لایه‌ی پنجم (۰/۰۰۷۵۵۹) و بیشترین مقدار آن برای استفاده از لایه‌ی هفتم (۰/۲۵۲۲) حاصل شده است. این نتایج نشان می‌دهد که همیشه میان نوع لایه و گرادیان رابطه‌ی مستقیمی وجود ندارد، زیرا شبکه در درجه‌ی اول سعی دارد عملکرد را بهینه کند و مقدار گرادیان بسته به بهترین عملکرد تعیین می‌شود.



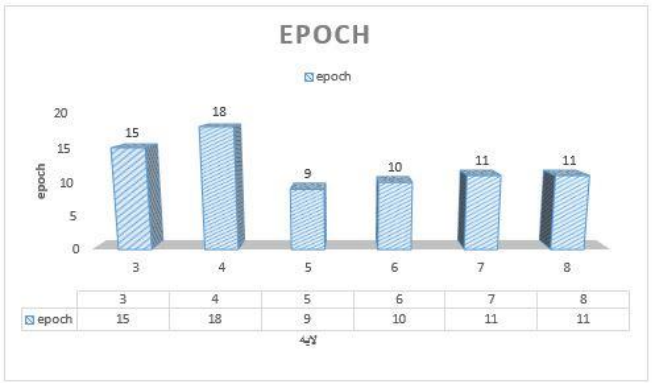
نمودار ۴-۳. رابطه‌ی لایه‌ی استفاده‌شده به عنوان لایه‌ی پنهان و گرادیان

نمودار ۴-۲ رابطه‌ی لایه‌ی استفاده‌شده به عنوان لایه‌ی پنهان و عملکرد را نشان می‌دهد، در این روند منظم‌تری را شاهدیم، با استفاده از لایه‌ی پنجم بهترین عملکرد حاصل شده است که مقدار آن قابل قبول است.

شبکه‌ی عصبی برای پیدا کردن نقطه‌ی مینیمم تابع performance ابتدا یک نقطه از آن را به‌صورت تصادفی انتخاب می‌کند، سپس با تکنیک‌های خاص به سراغ نقطه‌ی دوم می‌رود که مقدار این نقطه از نقطه‌ی اول کمتر است و همین روند را حفظ می‌کند. شبکه در هر تکرار از نقطه‌ای به نقطه‌ی دیگر می‌رود. رابطه‌ی تعداد تکرارها و تعداد لایه‌ها نیز رابطه‌ای تجربی است.

نمودار ۴-۴. رابطه‌ی لایه‌ی استفاده‌شده به عنوان لایه‌ی پنهان و عملکرد

نمودار ۴-۳ رابطه‌ی میان لایه‌ی استفاده‌شده به عنوان لایه‌ی پنهان و تعداد تکرارها یا دوره‌های زمانی لازم جهت دستیابی به بهترین عملکرد را نشان می‌دهد. استفاده از لایه‌ی پنجم منجر به بهترین عملکرد می‌شود و اما مزایای دیگری دارد که باز هم آن را به بهترین حالت ممکن تبدیل می‌کند، در شبکه‌ی کنونی با تعداد کم تکرارها نیز به نتیجه‌ی مناسبی دست خواهیم یافت، اما اگر در شبکه‌های مشابه نیاز به افزایش تعداد تکرارها یا دوره‌ها باشد، مسئله‌ی مدت زمان لازم برای آموزش شبکه اهمیت فراوانی می‌یابد. با تغییر لایه‌ی مورد استفاده از ۴ به ۵ لایه، تعداد تکرارها به نصف رسیده است، این نشان از توجیه‌پذیری زمانی استفاده از لایه‌ی پنجم به جای لایه-های دیگر است.

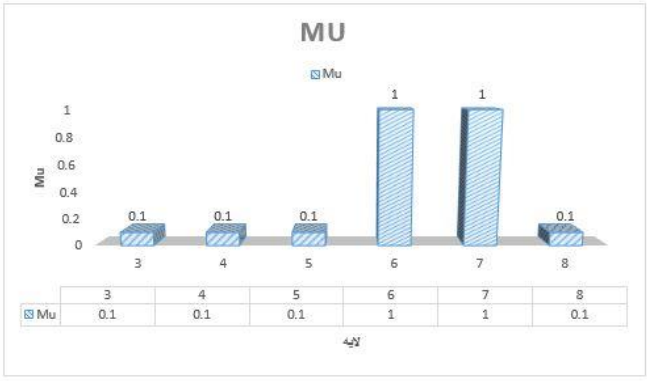


نمودار ۴-۵. رابطه‌ی لایه‌ی استفاده‌شده به عنوان لایه‌ی پنهان و تعداد تکرارها

www.SID.ir

Archive of SID

MU یک پارامتر میان برنامه‌ای است و رسیدن آن به مقدار عددی ۱ در استفاده از لایه‌های ۴ و ۵ نشان از بهینه بودن این لایه‌ها دارد؛ اما در تحلیل تعداد تکرارها دیدیم که استفاده از لایه‌ی چهارم منجر به تعداد دو برابری تکرارها نسبت به استفاده از لایه‌ی پنجم می‌شود. پس تا به اینجا نیز استفاده از لایه‌ی پنجم قابل توجیه است.



نمودار ۴-۶. رابطه‌ی لایه‌ی استفاده‌شده به عنوان لایه‌ی پنهان و MU

طبق آزمایش‌های انجام شده و همان گونه که در نمودارهای بالا مشاهده می‌کنید در شبکه عصبی طراحی شده مزایایی استفاده از لایه پنجم را نسبت به دیگر انتخاب‌ها نشان می‌دهد.

۵. جمع‌بندی و کارهای آتی

در این پژوهش تا جای که می‌دانیم برای اولین بار پیش‌بینی افراد خبره در شبکه‌های اجتماعی انجام شده است. علیرغم کلیه‌ی نارسایی‌ها و فرض‌های ساده کننده که در پژوهش پیش‌رو استفاده نمودیم، نتایج حاصله از استحکام و دقت بالایی برخوردارند، اجرای متوالی شبیه‌سازی‌ها و مشاهده‌ی نتایج باوجود تغییرات فراوان در ساختمان و چیدمان شبکه گویای این مدعاست. شبکه‌های اجتماعی با سرعت روز افزونی به پیشرفت خود ادامه می‌دهند و مطمئناً میزان نفوذ آن‌ها در حوزه‌های مختلف افزایش قابل ملاحظه‌ای خواهد داشت، گاهاً بررسی چنین شبکه‌هایی نیازمند حجم عظیمی اطلاعات است که در پایگاه‌های داده موجود بوده و دسترسی بدان‌ها، نرم‌افزارهایی تخصصی را می‌طلبد که به تبع بر پایه‌ی چنین پژوهش‌هایی استوارند. در این بررسی‌ها حوزه‌ی وسیعی از داده‌ها مورد استفاده قرار گرفته و اطلاعات فراوانی تولید می‌شود که مدیریت و تحلیل این اطلاعات نیز به نوبه‌ی خود، موضوع جالبی است. علیرغم تحقیقات گسترده و بسیاری که درزمینه شبکه‌های اجتماعی و پیش‌بینی افراد خبره برای این شبکه‌ها انجام شده است اما هنوز مسائل بسیاری وجود دارد که می‌تواند موضوعات تحقیقی بسیار خوبی در این زمینه باشد در این پژوهش مسئله پیش‌بینی افراد خبره به کمک الگوریتم‌های داد کاوی و شبکه‌های عصبی بررسی شد. علاوه بر این الگوریتم، می‌توان از الگوریتم فرابتکاری دیگر مانند الگوریتم ژنتیک برای این چالش در شبکه‌های اجتماعی نیز استفاده کرد. مهم‌ترین چالش در این کار، هماهنگ کردن مسئله جهت حداقل نمودن زمان و افزایش افراد خبره خواهد بود.

[1] Kardan, A. A. Omidvar, and F. Farahmandnia, Expert finding on social network with link analysis approach. 19th Iranian Conference on Electrical Engineering (ICEE), 2011: p. 1-6.

[2]Wang, D. T. Abdelzaher, and L. Kaplan, Understanding the social network. Social Sensing, 2015: p. 129-151.

[3] Bozzon, A. et al. Choosing the Right Crowd: Expert Finding in Social Networks. 2013

[4] C. S. Campbell, P. P. Maglio, A. Cozzi, B. Dom. Expertise Identification Using EmailCommunications. In Proc. of CIKM'03. 2003, pp.528-531

[5] J. Zhang, J. Tang, and J. Li, Expert FInding in a social network. Advances in Databases: Concepts, Systems and Applications, 2007. 4443(Lecture Notes in Computer Science): p. 1066-1069.

[6]Rafiei, M.Kardan, A. A (2015) A novel method for expert finding in online communities based on concept map and PageRank, Human-centric Computing and Information Sciences 5:10 DOI 10.1186/s13673-015-0030-5.

[7]Maron, M.E. S. Curry, and P. Thompson, An inductive search system: theory, design, and implementation. IEEE Trans. on Systems, Man and Cybernetics 1986. 16(1): p. 21-28.

[8]Balog, K. L. Azzopardi, and M.d. Rijke, Formal models for expert FInding in enterprise corpora. 2006: p. 43-55.

[9]Aslay, C. et al. Competition-Based Networks for Expert Finding. acm.org. SIGIR'13, 2013: p. ACM 978-1-4503-2034-4/13/07

[10] Riahi, F. Zolaktaf, Z. Shafiei, M. (2012) Finding Expert Users in Community Question Answering, WWW 2012 – CQA'12 Workshop, April 16–20, 2012, Lyon, France. ACM 978-1-4503-1230-1/12/04

[11]Seltzer, T. & Mitrook, M. (2007). The dialogic potential of weblogs in relationship building. Public Relations Review, 33(2), 227-229.

[12] Bortree, D. S. & Seltzer, T. (2009). Dialogic strategies and outcomes: An analysis of environmental advocacy groups' Facebook profiles. Public Relations Review, 35(3), 317–319.

[13] P. Ferragina and U. Scaiella. Tagme: on-the-y annotation of short text fragments (by Wikipedia entities). In Proceedings of the 19th ACM international conference on Information and knowledge management, CIKM '10, pages 1625{1628, New York, NY, USA, 2010. ACM.

www.SID.ir