دانشکده کامپیوتر و فناوری اطلاعات

پایان‌نامه کارشناسی ارشد هوش مصنوعی و رباتیک

شناسایی افراد مؤثر در شبکه اجتماعی توییتر

نگارنده: علی جعفری بلالمی

استاد راهنما

دکتر حمید حسن‌پور

استاد مشاور

دکتر باقر رحیم‌پور کامی

بهمن ۱۳۹۹

شماره:
تاریخ:

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد آقای علی جعفری بلالمی با شماره دانشجویی ۹۶۰۴۸۴۴ رشته مهندسی کامپیوتر گرایش هوش مصنوعی تحت عنوان شناسایی افراد مؤثر در شبکه اجتماعی توییتر که در تاریخ ۱۳۹۹/۱۱/۲۹ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می گردد:

الف) درجه عالی: نمره ۲۰-۱۹ ☐ ب) درجه خیلی خوب: نمره ۱۸/۹۹ – ۱۸ ☑

ج) درجه خوب: نمره۱۷/۹۹– ۱۶ ☐ د) درجه متوسط: نمره ۱۵/۹۹-۱۴ ☐

ه) کمتر از ۱۴ غیر قابل قبول و نیاز به دفاع مجدد دارد ☐

نوع تحقیق: **نظری** ☑ **عملی** ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول:	دکتر حمید حسن‌پور	استاد	
۲- استادراهنمای دوم			
۳- استاد مشاور:	دکتر باقر رحیم‌پور کامی		
۴- نماینده تحصیلات تکمیلی:	مهندس محسن فرهادی	مربی	
۵- استاد ممتحن اول:	دکتر حسین مرشدلو	استادیار	
۶ـ استاد ممتحن دوم:	دکتر مرضیه رحیمی	استادیار	

نام و نام خانوادگی رئیس دانشکده:

تاریخ و امضاء و مهر دانشکده:

تقدیم به خانواده عزیزم

فرشتگان مهربانی که لحظات ناب باور بودن، لذت و غرور دانستن، جسارت خواستن، عظمت رسیدن و تمام تجربه‌های یکتا و زیبای زندگیم، مدیون حضور سبز آن‌هاست.

سپاس‌گزاری

از استاد گرامیم جناب آقای دکتر **حمید حسن‌پور** بسیار سپاسگزارم چرا که بدون راهنمایی‌های ایشان تأمین این پایان‌نامه بسیار مشکل می‌نمود.

از جناب آقای دکتر **باقر رحیم‌پور کامی** سپاسگزارم به دلیل یاری‌ها و راهنمایی‌هایشان که بسیاری از سختی‌ها را برایم آسان‌تر نمودند.

علی جعفری بلالمی
بهمن ۱۳۹۹

تعهد نامه

اینجانب **علی جعفری بلالمی** دانشجوی کارشناسی ارشد رشته **هوش مصنوعی** دانشکده **کامپیوتر و فناوری اطلاعات** دانشگاه شاهرود، نویسنده پایان‌نامه با عنوان **شناسایی افراد مؤثر در شبکه اجتماعی توییتر** ، تحت راهنمایی **آقای دکتر حمید حسن‌پور** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است.
- در استفاده از نتایج پژوهش‌های دیگر پژوهش‌گران، به مرجع مورد استفاده استناد شده است.
- مطالب این پایان‌نامه، تا کنون توسط خود، یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ‌جا ارایه نشده است.
- حقوق معنوی این اثر، به دانشگاه صنعتی شاهرود تعلق دارد، و مقالات مستخرج با نام ‘‘ دانشگاه صنعتی شاهرود’’ یا “ Shahrood University of Technology ” به چاپ خواهد رسید.
- حقوق معنوی تمام افرادی که در به‌دست آوردن نتایج اصلی پایان‌نامه تاثیرگذار بوده‌اند، در مقالات مستخرج از پایان‌نامه رعایت می‌گردد.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که از موجود زنده (یا بافت‌های آنها) استفاده شده است، ضوابط و اصول اخلاقی رعایت شده است.
- در تمام مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته (یا استفاده شده است)، اصل رازداری و اصول اخلاق انسانی رعایت شده است.

علی جعفری بلالمی
بهمن ۱۳۹۹

مالکیت نتایج و حق نشر

- تمام حقوق معنوی این اثر و محصولات آن (مقالات مستخرج، کتاب، برنامه‌های رایانه‌ای، نرم‌افزارها و تجهیزات ساخته شده) متعلق به دانشگاه صنعتی شاهرود می‌باشد. این مطلب باید به نحو مقتضی، در تولیدات علمی مربوطه ذکر شود.
- استفاده از اطلاعات و نتایج موجود در این پایان‌نامه بدون ذکر منبع مجاز نمی‌باشد.

چکیده

امروزه شناسایی افراد تأثیرگذار در شبکه‌های اجتماعی به یکی از موضوعات داغ و پر طرف‌دار تبدیل شده‌است. محبوبیت این موضوع مربوط به دلایل مختلفی از جمله کاهش هزینه و هدفمند کردن تبلیغات است. با این حال چالش‌های زیادی بر سر راه محققین در این زمینه وجود دارد. از جمله این چالش‌ها نیز می‌توان به حجم زیاد کاربران در شبکه‌های اجتماعی و فعالیت‌های آن‌ها، نحوه‌ی تعریف تأثیر اجتماعی، انتخاب مؤلفه‌های مناسب برای اندازه‌گیری میزان تأثیر و دسترسی محدود به اطلاعات بدلیل حفظ حریم شخصی کاربران در شبکه‌های اجتماعی نام برد.

هدف این پایان‌نامه یافتن افراد تأثیرگذار در شبکه اجتماعی توییتر با ارائه‌ی تعریفی مناسب برای مفهوم تأثیر و همچنین انتخاب مناسب‌ترین مؤلفه‌ها برای اندازه‌گیری میزان تأثیر می‌باشد.

در روش پیشنهادی برای محاسبه‌ی میزان تأثیر هر فرد ابتدا میزان تأثیر هر توییت محاسبه می‌شود. میزان تأثیر هر توییت در این روش برابر است با تعداد افرادی که بصورت بالقوه می‌توانند آن توییت را مشاهده کنند. برای این کار نیاز است مخاطبین آشکار و پنهان هر توییت شناسایی شوند.

برای پیدا کردن توییت‌هایی که احتمالا تحت تأثیر توییت فردی دیگر منتشر می‌شوند باید بدنبال توییت‌هایی با محتوای مشابه بود. برای یافتن توییت‌های مشابه از الگوریتم LSH استفاده شده تا با توجه به حجم بالای داده‌ها، میزان محاسبات و هزینه زمانی کاهش یابد. طبق روش پیشنهادی میزان تأثیر هرکاربر برابر با میانگین تأثیر توییت‌های آن کاربر است. اما در مقام مقایسه، برای رتبه‌بندی کاربران بر اساس مقدار تأثیر در صورتی که میانگین تأثیر توییت‌هایشان برابر باشد، مقدار تأثیر کاربری که واریانس تأثیر توییت‌هایش کمتر باشد، بیشتر است. همچنین در صورتی که هر دو مقدار میانگین و واریانس برای دو کاربر برابر باشد، کاربری که تعداد توییت‌هایش بیشتر باشد، تأثیر بیشتری خواهد داشت.

با توجه به روش ارائه شده و مؤلفه‌های مورد نیاز و همچنین وجود نداشتن پایگاه داده مناسب و عدم امکان جمع‌آوری داده، از اشتراک دو پایگاه داده ArchiveTeam و Twitter Social Graph داده‌های مناسب و مرتبط استخراج شده و مورد استفاده قرار گرفت. در نهایت طبق نتایج بدست آمده، روش پیشنهادی با استفاده از مؤلفه‌هایی کارا تر از کارهای پیشین به خوبی قادر به محاسبه‌ی میزان دقیق‌تر و واقعی تأثیر کاربران می‌باشد.

کلمات کلیدی: شبکه‌های اجتماعی، افراد تأثیرگذار، LSH، نفوذ اجتماعی، توییتر

لیست مقالات مستخرج از پایان‌نامه

- Jafari. A., Hassanpour. H, Rahimpour. B., "Discovering influential users on Twitter" (in progress)

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱

مقدمه

با پیشرفت چشمگیر تکنولوژی در دهه‌های اخیر، در دسترس بودن کامپیوترها، ورود تلفن‌های همراه هوشمند به زندگی انسان، وجود اینترنت و سهولت برقراری ارتباط سریع برای اکثریت افراد در سراسر دنیا، جاذبه و محبوبیت زیادی برای شبکه‌های اجتماعی مجازی بوجود آمده است. شبکه‌های اجتماعی مانند Facebook، Twitter، Instagram، sina weibo و غیره، کاربران وب در سراسر دنیا را به شکل اجتناب ناپذیری به یکدیگر متصل کرده‌اند. استقبال حداکثری مردم از این شبکه‌ها نیز پژوهشگران در حوزه‌های مختلف را جذب تحلیل این شبکه‌ها کرده است.

۱.۱ مروری بر شبکه‌های اجتماعی

بر اساس زبان ریاضی، شبکه‌های اجتماعی به شکل یک گراف درنظر گرفته می‌شود، به صورتی که موجودیت‌ها به شکل گره‌های گراف و روابط بین آن‌ها نیز به شکل یال‌های متناظر تصویر می‌شود.

در شبکه‌های اجتماعی مجازی، هریک از کاربران را به شکل یک گره از گراف آن شبکه در نظر گرفته و یال‌های شبکه نیز روابط بین کاربران را که می‌تواند نشانگر روابط دوستی، دنبال‌کردن، ارسال پیام یا غیره باشد، نشان می‌دهد.

به عقیده اگاروال شبکه اجتماعی [۱] نوعی شبکه است که منعکس کننده ساختار اجتماعی

گره‌های آن و وابستگی متقابل آن‌ها مانند دوستی با مردم، همکاری مؤلفان، جستجوگرها و همکاری بین طرف‌های مختلف است. با یک شبکه اجتماعی می‌توان به عنوان یک شبکه‌ی پیچیده رفتار کرد که از افراد جامعه و روابط آن‌ها در بین افراد تشکیل شده است. حجم این شبکه‌ها بسیار زیاد است. این نوع ساختار پیچیده اجتماعی، عمدتا نقش مهمی در حجم انتشار اطلاعات دارد.

توسعه چشمگیر شبکه‌های اجتماعی، افراد بیشتری را به انجام فعالیت‌های مختلف اجتماعی جذب کرده است، از جمله آن‌ها می‌توان به گفت‌وگوی آنلاین، ایجاد دوستی از طریق اسکایپ، انتشار محتوا یا اظهار نظر از طریق میکروبلاگ‌ها، مرور پروفایل از طریق اینترنت، پیام رسانی از طریق تلفن‌های هوشمند و خرید آنلاین از طریق بستر تجارت الکترونیکی اشاره کرد.

۱.۱.۱ انواع شبکه‌های اجتماعی از لحاظ توپولوژی

۱. **شبکه‌های اجتماعی آنلاین (OSN)** [1] :

شبکه‌های اجتماعی آنلاین، جوامع مجازی هستند که به افراد اجازه اتصال و تعامل با یکدیگر در یک موضوع خاص یا صرفا فعالیت به صورت آنلاین را می‌دهد. با ظهور وب (WWW) در دهه ۱۹۹۰ و توسعه سریع فن‌آوری‌های مختلف ICT، شبکه‌های اجتماعی آنلاین به صورت نمایی رشد کرده‌اند. به همین دلیل در سال‌های اخیر، شبکه‌های اجتماعی آنلاین به طور چشمگیری در سراسر جهان محبوبیت پیدا کرده‌اند. به عنوان مثال براساس داده‌های جمع‌آوری‌شده در ماه اکتبر ۲۰۲۰ [۲]، فیس بوک ۲/۶ میلیارد کاربر در هر ماه داشته است. علاوه بر موارد اشاره شده، برای نمونه تعداد کاربران شبکه‌های اجتماعی محبوب نیز در جدول ۱.۱ آورده شده است. رشد سریع شبکه‌های اجتماعی آنلاین تعداد زیادی از محققان را به اکتشاف و بررسی این خدمات عمومی فراگیر، جلب کرده‌است.

۲. **شبکه‌های اجتماعی سیار (MSN)** [2] :

با توسعه چشمگیر فناوری‌های شبکه بی‌سیم (به عنوان مثال، شبکه تلفن همراه، WiFi و بلوتوث)، برنامه‌های شبکه آنلاین اجتماعی مبتنی بر وب منجر به تغییر شبکه‌های آنلاین به شبکه‌های سیار شده است.

طبق توپولوژی شبکه، MSNها [۳۶] را می‌توان به سه طبقه تقسیم‌کرد [۴]. یکی از آن‌ها MSN متمرکز است که یکی از بسطهای OSN است. کاربران با استفاده از برنامه‌های خاص تلفن همراه یا یک مرورگر موبایل، قادر به اتصال به یک سرور متمرکز برای اشتراک و تبادل اطلاعات هستند. مورد دوم، MSNهای توزیع‌شده است که به کاربران امکان برقراری ارتباط و دسترسی به اطلاعات اجتماعی را از طریق شیوه‌ای

[1] Online Social Networks
[2] Mobile Social Networks

جدول ۱.۱: اطلاعات مقدماتی برخی از شبکه‌های اجتماعی مشهور

شبکه اجتماعی	تاسیس	کشور	کاربران فعال ماهانه	کاربرد
Facebook	۲۰۰۴	آمریکا	۲/۶ میلیارد	شبکه‌اجتماعی
WhatsApp	۲۰۱۰	آمریکا	۱/۶ میلیارد	پیام رسان
QQ	۱۹۹۹	چین	۷۳۱ میلیون	پیام رسان
Qzone	۲۰۰۵	چین	۵۱۷ میلیون	شبکه‌اجتماعی
WeChat	۲۰۱۱	چین	۱/۱ میلیارد	پیام رسان
Instagram	۲۰۱۰	آمریکا	۱ میلیارد	اشتراک رسانه
Twitter	۲۰۰۶	آمریکا	۳۴۰ میلیون	میکروبلاگینگ
weibo Sina	۲۰۰۹	چین	۴۹۷ میلیون	میکروبلاگینگ
Snapchat	۲۰۱۱	آمریکا	۳۸۲ میلیون	اشتراک رسانه
Pinterest	۲۰۱۰	آمریکا	۳۲۲ میلیون	اشتراک رسانه
Reddit	۲۰۰۵	آمریکا	۴۳۰ میلیون	اخبار اجتماعی
YouTube	۲۰۰۵	آمریکا	۲ میلیارد	اشتراک رسانه

موقت و بدون نیاز به هیچ نوع زیرساختی می‌دهد. آخرین مورد، MSNهای ترکیبی است که تعمیم‌یافته‌ی MSNهای متمرکز و MSNهای توزیع‌شده است. براین اساس کاربران می‌توانند با استفاده از زیرساخت‌های سلولی به اطلاعاتی دسترسی داشته‌باشند یا آن‌ها را به اشتراک بگذارند و یا ممکن است اطلاعات را توسط ad-hoc به اشتراک بگذارند.

به دلیل مزایای غیرقابل انکار این شبکه‌ها، مانند استقرار آسان و ارزان و ارتباط برقرار شده بین موقعیت فیزیکی کاربران و تعاملات اجتماعی آن‌ها، برنامه‌های اجتماعی موبایل توجه بسیاری را به خود جلب کرده‌اند. تنوع برنامه‌های مبتنی بر شبکه‌های اجتماعی موبایل بسیار چشمگیر است، محبوب ترین دسته برنامه‌ها شامل شبکه‌های اجتماعی (مانند Facebook)، پیام رسانی فوری (مانند WhatsApp)، اشتراک رسانه‌ای (مانند Instagram) و غیره است.

۲.۱.۱ انواع شبکه‌های اجتماعی از لحاظ عملکرد

۱. **اخبار اجتماعی**[1]:

یک وب‌سایت اخبار اجتماعی، دارای پست‌های ارسال شده توسط کاربران است که براساس محبوبیت رتبه‌بندی می‌شوند [۵]. کاربران می‌توانند در مورد این پست‌ها نظر

[1] Social news

دهند و ممکن است این نظرات مانند Reddit رتبه‌بندی شوند.

۲. **پیام رسانی فوری(IM)**[۱]:

نوعی پلتفرم چت آنلاین است که انتقال متن را درلحظه از طریق اینترنت فراهم می‌کند. یک پیام رسان مبتنی بر شبکه LAN (مانند Tencent QQ) با روشی مشابه از طریق شبکه محلی فعالیت می‌کند.

۳. **میکروبلاگینگ**[۲]:

این نوع شبکه به کاربران امکان می‌دهد عناصر کوچکی از محتوا مانند جملات کوتاه، تصاویر منفرد یا پیوندهای ویدئویی را ردوبدل کنند [۶، ۷] ، بعنوان مثال توییتر [۸] یک پلتفرم میکروبلاگینگ است که به کاربران ثبت شده اجازه می‌دهد متنی کوتاه را از راه‌های مختلفی از جمله وب، پیام کوتاه و IM منتشر کنند.

۴. **نشانه گذاری اجتماعی**[۳]:

این شبکه یک سرویس آنلاین متمرکز است که اجازه اضافه کردن، ویرایش، حاشیه‌نویسی و به اشتراک گذاری اسناد وب را به کاربران می‌دهد، مانند StumbleUpon [۹] که یک موتور کشف و تبلیغات بود که محتوای وب را به کاربران توصیه می‌کرد. این سایت در نوامبر ۲۰۰۱ تأسیس شده و در ژوئن ۲۰۱۸ بسته شد.

۵. **فروم اینترنتی**[۴]:

فروم یک سایت گفت‌وگوي آنلاین است که مردم می‌توانند در آن مکالمات را به شکل پست‌هایی ارسال کنند [۱۰].فروم‌ها غالبا موضوع‌های مشخصی دارند به همین خاطر با اتاق‌های گپ تفاوت اساسی دارند زیرا پیام‌ها اغلب متن‌هایی طولانی هستند و بایستی موقتاً بایگانی شوند. به عنوان مثال، Douban[۱۱]یک وب سایت چینی است که به کاربران اجازه می‌دهد تا اطلاعات را ثبت کنند و محتوای مرتبط با فیلم‌ها، کتاب‌ها، و حوادث و فعالیت‌های اخیر در شهرها را ارسال کنند.

۶. **وب سایت اشتراک رسانه**[۵]:

این شبکه به کاربران امکان می‌دهد اطلاعات مختلف مانند فیلم‌ها، عکس‌ها و موسیقی را بارگذاری یا مشاهده‌کنند، همچنین در باره‌ی آن‌ها نظر دهند و به اشتراک بگذارند. به عنوان مثال، YouTube [۱۲] یک وب سایت اشتراک ویدیو است که به کاربران امکان

[1] Instant Messaging
[2] Microblogging
[3] social bookmarking
[4] Internet Forum
[5] Media Sharing Website

می‌دهد فیلم‌ها را بارگذاری، مشاهده، رتبه‌بندی، یا اشتراک گذاری کنند و یا نظر خود را در باره‌ی آن‌هابیان کنند.

۷. **شبکه سازی اجتماعی**[۱]:

این شبکه به کاربران، این امکان را می‌دهد که ایده‌ها، پست‌ها، تصاویر، فعالیت‌ها، رویدادها و علایق خود را با افراد موجود در شبکه‌هایی مانند فیس‌بوک به اشتراک بگذارند [۱۳].

۲.۱ تحلیل شبکه‌های اجتماعی

تحلیل شبکه اجتماعی، هم یک حوزه تحقیق بین‌رشته‌ای است و هم یک تکنیک تحلیلی. این فرایند به این سؤال پاسخ می‌دهد که ساختار روابط بین موجودات اجتماعی چگونه است و این ساختار چه تأثیری بر سایر پدیده‌های اجتماعی دارد. بنابراین اصلی‌ترین هدف این روش، کشف و تفسیر الگوی پیوندهای اجتماعی بین کنشگران است.

برای تحلیل شبکه‌های اجتماعی، پارامترها و شاخص‌های متعددی طراحی شده و مورد استفاده قرار می‌گیرد. میزان تعامل هر گره[۲] با گره‌های دیگر، تفاوت یا تشابه توزیع جغرافیایی گره‌ها و توزیع دیجیتال آن‌ها، عمق و شدت نفوذ اثر هر رفتار گره بر روی گره‌های دیگر، قرار گرفتن هر گره در مرکز یا میزان فاصله گرفتن آن از مرکز شبکه، تعامل یکسویه یا دوسویه‌ی گره با سایر گره‌ها، تنوع اطلاعات و ارتباطات بین گره‌ها و آنتروپی موجود در شبکه اجتماعی، از جمله صدها پارامتر و معیاری هستند که در تحلیل شبکه‌های اجتماعی مورد توجه قرار می‌گیرند.با استفاده از تحلیل شبکه، می‌توان مجموعه‌های پیچیده‌ای از روابط را به مثابه نقشه‌هایی (گراف یا نگاره‌های گروهی) از سمبل‌های متصل تجسم کرد و سنجه‌های دقیق اندازه شکل و تراکم شبکه را به مثابه یک کل و موقعیت هر عنصر را درون آن محاسبه نمود. با توجه به آن چه اشاره گردید می‌توان گفت که تحلیل شبکه‌های اجتماعی کمک می‌کند تا الگوهای موجود درون مجموعه‌ی نهادهای مرتبط را که شامل مردم هستند، تجسم و بررسی کنیم.

بدین ترتیب می‌توان افراد تأثیرگذار[۳] در جامعه را به منظورهای مختلف شناسایی کرد و یا با شناسایی گروه‌ها[۴]، گروه‌هایی که با منظور خاص و یا بدون هماهنگی قبلی و به واسطه‌ی فعالیت‌های کاربران شکل گرفته‌اند، به درستی مورد تجزیه و تحلیل قرار داد.

با تجزیه‌وتحلیل و استخراج اطلاعات از شبکه‌های اجتماعی، می‌توانیم اطلاعاتی را درباره نظرات مردم در مورد یک محصول خاص جمع آوری کنیم. تجزیه‌وتحلیل این گونه نظرات،

[1] Social Networking
[2] Node
[3] influencer
[4] Community Detection

ارزش آن را برای طراحی بازاریابی و تبلیغات نشان می‌دهد. از نمونه‌های معمول این نوع تحلیل‌ها می‌توان به بازاریابی ویروسی [۱۴، ۱۵]، یافتن وبلاگ نویسان تأثیرگذار [۱۶، ۱۷]، تبلیغات آنلاین [۱۸]، مراقبت‌های بهداشتی اجتماعی [۱۹، ۲۰]، یافتن متخصصان [۲۱، ۲۲]، توصیه‌های شخصی [۲۳]، شبکه‌های استنادی [۲۴، ۲۵]، شناسایی شایعات [۲۶]، تحلیل احساسات [۲۷] و غیره اشاره نمود. البته با توجه به ماهیت و موضوع مورد نظر، ممکن است یک تحقیق برای رسیدن به هدف خود نیازمند به استفاده از چند نمونه از این تحلیل‌ها باشد. بعنوان مثال برای تبلیغات و بازاریابی آنلاین، ممکن است «تحلیل احساسات» (برای پی‌بردن به نظر کاربران در مورد محصول یا خدمات) و یا استفاده از «یافتن متخصصان» و «یافتن وبلاگ نویسان تأثیرگذار» (برای هدفمند کردن تبلیغات)، بخشی از روش مورد نظر یک تحقیق باشد.

۱.۲.۱ تحلیل تأثیر اجتماعی

فرایند تحلیل تأثیر اجتماعی [۲۸] در حال تبدیل شدن به بخش مهمی از شبکه‌های اجتماعی است. با تجزیه‌وتحلیل نحوه تأثیرگذاری در میان کاربران و نحوه گسترش نفوذ بر اساس داده‌های بزرگ شبکه‌های اجتماعی [۲۹]، می‌توان مزایای زیر را بدست آورد:

۱. از نظر جامعه شناسی: درک رفتارهای اجتماعی افراد

۲. از نظر خدمات عمومی: ارائه مبنای نظری برای تصمیم گیری عمومی و راهنمایی افکار عمومی

۳. از نظر کشور: ارتقا امنیت ملی، ثبات اقتصادی، پیشرفت اقتصادی و غیره

در نتیجه، تحلیل تأثیر اجتماعی در شبکه‌های اجتماعی دارای اهمیت و ارزش کاربردی مهمی است. با این حال، با توجه به افزایش قابل‌توجه شبکه‌های اجتماعی در مقیاس و حجم، ما با چالش‌های متعددی در اندازه‌گیری نفوذ اجتماعی برای یک کاربر معین مواجه هستیم.[۲۸] موانع زیادی بر سر این راه وجود دارد که در ادامه به آن‌ها اشاره می‌کنیم. نخست این‌که، نفوذ اجتماعی عمدتا یک مفهوم انتزاعی است و ما تعریف ریاضی و مقیاس واحدی برای اندازه‌گیری آن نداریم. دوم، تصمیم گیری در مورد عوامل اصلی برای یک مورد خاص، جهت مدل سازی تأثیر اجتماعی بسیار دشوار است. سوم این‌که هیچ روش موثری برای ادغام مناسب عوامل مختلف برای اندازه‌گیری تأثیر وجود ندارد. در نهایت با این مساله مهم مواجه هستیم که، مشخص کردن رابطه علت و معلولی تأثیر اجتماعی و عدم قطعیت نفوذ اجتماعی بسیار دشوار است.

محققان بخش‌های مختلف علوم رایانه، علاوه بر تجزیه‌وتحلیل تأثیرات اجتماعی، تحقیقات گسترده‌ای را انجام می‌دهند تا کاربردهای شبکه‌های اجتماعی ممکن وکاربردی شود. امروزه، تحلیل تأثیرگذاری اجتماعی به موضوعی داغ در شبکه‌های اجتماعی تبدیل شده است. مقالات

زیادی [۲۳، ۳۰]، کتاب [۳۱]، نظرسنجی‌ها [۳۲] و آموزش‌های [۳۳] مربوط به تحلیل تأثیر اجتماعی وجود دارد. ما شاهد پیشرفت چشمگیر در این زمینه نوظهور هستیم که توسط برنامه‌های کاربردی عملی از طریق تحلیل تأثیر اجتماعی مورد استفاده قرار گرفته است. با این وجود ظهور شبکه‌های اجتماعی مشکلات بی‌سابقه بسیاری را بوجود می‌آورد که با ابزارها و تئوری‌های موجود قابل حل نیست. با توسعه قابل‌توجه شبکه‌های اجتماعی، به خصوص حجم فوق‌العاده از مجموعه داده‌های بزرگ موجود در شبکه‌های اجتماعی، سوالات فراوانی همچنان بی‌پاسخ مانده‌اند.

تحلیل تأثیر اجتماعی یک حوزه منظم است. همان‌طور که تحقیقات در مورد شبکه‌های اجتماعی هنوز در مراحل ابتدایی خود است، مطالعه تأثیر اجتماعی ارتباط interconnects با دیگر ویژگی‌های شبکه‌های اجتماعی، مانند ویژگی‌های نفوذ، معیارهای ارزیابی تأثیر، جمع‌آوری و پردازش داده‌های بزرگ شبکه‌های اجتماعی و انتخاب k گره برتر تأثیر گذار را نشان می‌دهد[۳۴]. بنابراین، بررسی تحلیل تأثیر اجتماعی در شبکه‌های اجتماعی در مقیاس بزرگ تحت محیط جهانی داده‌های بزرگ ضروری است. بر این اساس، ما قصد داریم که یک تصویر نسبتا کامل در مرکز تحلیل نفوذ اجتماعی ارایه کنیم.

ما باید تأثیر اجتماعی را عمیقا درک کنیم تا از این طریق بتوانیم در شبکه‌های اجتماعی بهتر عمل کنیم. تا کنون درک ما از تأثیر اجتماعی در بسیاری از جنبه‌ها کم‌عمق بوده است. از نظر نمایش، خود شبکه‌های اجتماعی را می‌توان به عنوان یک شبکه انتزاعی از گره‌های بهم مرتبط دانست. محققان عموما یک گراف یا ماتریس را برای توصیف شبکه‌ها استفاده می‌کنند. با این حال، در مقیاس بزرگ، یک گراف بزرگ (یا ماتریس)با میلیون‌ها گره و یال وجود خواهد داشت. تجزیه یا فشرده‌سازی چنین گراف‌های بزرگی چالش برانگیز است [۳۵]. علاوه بر این، توصیف ویژگی‌های دینامیکی شبکه‌های اجتماعی با استفاده از روش‌های سنتی نیز دشوار است.

ما بر این باوریم که این کار یک بررسی جامع و مفهومی بر روی تأثیر اجتماعی به ویژه در شبکه‌های اجتماعی است. ابتدا، به منظور ارایه درک بهتر از تحلیل تأثیر اجتماعی، مروری بر انواع شبکه‌های اجتماعی ارایه می‌دهیم. تحلیل تأثیر اجتماعی نیازمند اندازه‌گیری دقیق تأثیر اجتماعی است، به همین دلیل ما خواهیم کوشید تا آثار موجود در ارزیابی تأثیر اجتماعی را خلاصه و تحلیل کنیم. بعد از آن، تعدادی از تحقیقات مرتبط را در مورد الگوریتم بیشینه‌سازی انتشار در شبکه‌های اجتماعی دسته‌بندی کرده و مقایسه می‌کنیم. سپس مسائل باز یافته از تحلیل تأثیر اجتماعی را شناسایی و فرصت‌های هیجان‌انگیز در شاخه تحقیقاتی نوظهور را ارایه می‌دهیم. هدف ما هموار کردن یک بنیان محکم برای جوامع مختلف، برای بررسی بیشتر این موضوع امیدبخش است.

تحلیل شبکه‌های اجتماعی (Social Network Analysis) که گاهی به اختصار به آن SNA هم گفته می‌شود به معنای فرایند بررسی و ارزیابی ساختارهای یک شبکه اجتماعی به عنوان یک گراف از ابزارها یا انسان‌هاست که با خطوط ارتباطی به یکدیگر متصل هستند.

این خطوط ارتباطی می‌تواند رابطه دوستی، انتقال ویروس بیماری، رابطه ارسال و دریافت پیام و پیامک، تبادل نظر در زمینه رای دادن در یک انتخابات، رابطه ارسال و دریافت بسته‌های اطلاعاتی و یا هر نوع رابطه دیگری باشد.

همان‌طور که در تعریف شبکه‌های اجتماعی اشاره گردید، تحلیل شبکه‌های اجتماعی با استفاده از دانش ریاضی و نظریه گراف‌ها، شکلی بسیار علمی و ساختاریافته به خود گرفته است و در حال حاضر تحلیل شبکه‌های اجتماعی از جمله گران‌ترین و پرسودترین تخصص‌های جوامع بشری محسوب می‌شود.

پیشنهاد دوستان جدید در فیس‌بوک و یا پیشنهاد تصاویر توسط اینستاگرام که ما به صورت روزانه با آن‌ها سر و کار داریم، از جمله خروجی الگوریتم‌های تحلیل شبکه‌های اجتماعی است.

دو نمونه از موضوع‌هایی که در زمینه‌ی تحلیل شبکه‌های اجتماعی مطرح بوده و برای ما دارای اهمیت هستند، «پیش‌بینی تأثیرگذاری خبر در شبکه‌اجتماعی قبل از انتشار» و «پیدا کردن افراد با نفوذ برای گسترش بیشتر خبر در شبکه» می‌باشد. پیدا کردن چنین افرادی در شبکه‌های اجتماعی می‌تواند کاربردهای زیادی داشته‌باشد. به‌عنوان مثال اگر شرکتی بخواهد بداند از چه طریقی محصولش را تبلیغ کند می‌تواند از این طریق به نتیجه مطلوب برسد.

بنابراین هدف ما در این مرحله، بررسی روش‌های متفاوت انجام این کار است. به این منظور تعدادی از مقالاتی که در این زمینه منتشر شده‌اند را مطالعه و روش‌های انجام کار پیشنهادی آن‌ها را مورد بررسی قرار می‌دهیم.

۲.۲.۱ درک تأثیر اجتماعی

در این بخش، ما تعاریف شبکه‌های اجتماعی و چارچوب تحلیل تأثیر اجتماعی، توصیف کاربردهای کلاسیک در تأثیر اجتماعی و مدل‌های نفوذ را ارائه می‌دهیم.

تعاریف نفوذ اجتماعی:

در این بخش، معنای تأثیر اجتماعی را توضیح خواهیم داد. ما تأثیر اجتماعی را به عنوان سطح عدم اطمینان تفسیر می‌کنیم. درک اساسی تأثیر اجتماعی به شرح زیر خلاصه می‌شود:

- نفوذ اجتماعی رابطه‌ای است که بین دو نهاد برای یک عمل خاص برقرار شده است. به طور خاص، یک موجودیت برای انجام عملی موجودیت دیگر را تحت تأثیر قرار می‌دهد. معمولاً به نهاد اول تأثیرگذار و به نهاد دوم تأثیرپذیر گفته می‌شود.
- از سوی دیگر تأثیر اجتماعی تابعی از عدم اطمینان است. به طور خاص، اگر اینفلوئنسر معتقد باشد که تأثیرپذیر عمل را به طور حتم انجام می‌دهد، اینفلوئنسر برای انجام عمل کاملاً «تأثیرگذار» است و هیچ شکی وجود ندارد. اگر اینفلوئنسر معتقد است که تأثیرپذیر مطمئناً عملی را انجام نمی‌دهد، اینفلوئنسر بر تأثیرپذیر برای عدم انجام عمل

«تأثیر نمی‌گذارد»، پس عدم قطعیت نیز وجود ندارد. اگر اینفلوئنسر هیچ تصوری از اینکه آیا تأثیرپذیر عمل را انجام می‌دهد یا خیر، ندارد، اینفلوئنسر بر آن تأثیرگذار نیست. در این حالت، اینفلوئنسر بیشترین عدم اطمینان را دارد [۳۶].

- سطح تأثیر اجتماعی را می‌توان با یک عدد حقیقی پیوسته اندازه‌گیری کرد، که به ارزش تأثیر اجتماعی اشاره دارد. نفوذ اجتماعی نیز می‌تواند با عدم قطعیت (به عنوان مثال، قوی‌تر، قوی، ضعیف، ضعیف‌تر و غیره) نمایش داده شود [۳۶].
- ممکن است تأثیرگذار مقادیر تأثیر اجتماعی مختلفی را برای یک عمل روی یک فرد (تأثیر پذیر) را اعمال کند. تأثیرات اجتماعی لزوما متقارن (دو طرفه) نیستند. در صورتی که A و B دو شخص باشند، این که شخص بر A شخص B تأثیر می‌گذارد لروما به این معنی نیست که شخص B هم بر شخص A تأثیر می‌گذارد.

تأثیر اجتماعی[۳۷] به صورت زیر تعریف می‌شود: با توجه به دو فرد u و v در یک شبکه اجتماعی، u قدرت را به v اعمال می‌کند، یعنی u تأثیر تغییر عقیده v به روش مستقیم یا غیر مستقیم را دارد.

مسئله بیشینه‌سازی تأثیر

مسئله بیشینه‌سازی تأثیر با فرض داشتن گراف دنبال‌کننده‌ها در شبکه، به دنبال یافتن مجموعه‌ای از گره‌ها با شرایط خاص (مجموعه با کمترین تعداد گره‌ها، مجموعه با کم‌هزینه‌ترین گره‌ها) است که با استفاده از این مجموعه گره‌ها و یک مدل انتشار[۱] بتوان کل شبکه را تحت تأثیر قرار داد.

به هر وضعیت از شبکه در لحظه، یک تصویر[۲] گفته می‌شود که بعد از هر تصویر، بر اساس مدل انتشار، وضعیت بعدی شبکه ساخته می‌شود. مسئله‌ی یافتن تعدا کمینه از این افراد یک مسئله‌ی NP-Complete است و در زمان چندجمله‌ای قابل حل نمی‌باشد. برای این منظور معمولا از روش‌های حریصانه استفاده می‌شود که یک جواب تقریبی و نزدیک به جواب واقعی بدست بیاید.

مدل انتشار

مدل انتشار تعریف و یا فرضی از نحوه‌ی عملکرد گره‌های شبکه درنظر گرفته شده است. بعنوان مثال در ادامه یک نمونه ساده از مدل انتشار آمده است. فرض می‌کنیم با انتخاب کمترین تعداد گره و فعال کردن آن‌ها، کل شبکه نیز فعال شود. برای مثال، گراف جهت‌دار مانند شبکه‌ی اجتماعی توییتر یا اینستاگرام را در نظر می‌گیریم. در اینجا فرض بر آن است که رفتار

[1] Diffusion Model
[2] Snapshot

همه‌ی کاربران در شبکه شبیه به یکدیگر است. همه‌ی آن‌ها به این شکل عمل می‌کنند که اگر یک کاربر خبری را از دو نفر از افرادی که آن‌ها را دنبال می‌کند دریافت نماید، خود او نیز آن خبر را منتشر خواهد کرد. وقتی یک فرد خبری را منتشر کند، همه‌ی افرادی که این شخص را دنبال می‌کنند، این خبر را دریافت می‌کنند. این مدل رفتاری به این شکل است که کاربر ممکن است خبری را از یکی از افرادی که دنبال می‌کند دریافت کند ولی آن را منتشر نکند مگر اینکه آن خبر را از شخص دیگری که آن را دنبال می‌کند، دریافت کند. در این صورت شخص قانع می‌شود که خبر، مهم و به احتمال زیاد درست است و آن را ارسال می‌کند. به این فرضیات، مدل انتشار در شبکه‌ی اجتماعی می‌گویند.

چارچوب تحلیل تأثیر اجتماعی:

- **جمع آوری داده‌های بزرگ از شبکه‌های اجتماعی** : این یک مبنای بسیار مهم در تحلیل تأثیر اجتماعی است. با در دسترس بودن داده‌های بزرگ تولید شده توسط خرید آنلاین، تبلیغات، پیام رسانی فوری، ارتباطات سیار، جمع آوری داده‌های خام از منابع آنلاین و آفلاین برای ما بسیار راحت‌تر است.

- **پیش‌پردازش داده‌های بزرگ** : به منظور بهبود عملکرد و سهولت پردازش، حذف اطلاعات بی ربط در مورد تحلیل تأثیر اجتماعی ضروری است. برای محافظت از حریم خصوصی کاربران، لازم است اطلاعات حساس مربوط به محافظت از حریم خصوصی را از مجموعه‌ی داده‌های جمع آوری شده حذف کنیم.

- **انتخاب معیارهای ارزیابی** : استخراج مجموعه‌ای از معیارهای ارزیابی برای سنجش تأثیر اجتماعی هر کاربر بسیار مهم است. از آن‌جا که این معیارهای ارزیابی برای تعیین کمیت تأثیر اجتماعی هر کاربر و یافتن kـ تأثیرگذارترین گره‌ها مفید هستند. معیارهای ارزیابی دراین کار عموما شامل مرکزیت هر گره، فرکانس تعامل و غیره است.

- **اندازه‌گیری تأثیر اجتماعی** : با توجه به معیارهای ارزیابی استخراج شده، مدل ارزیابی و معادلات محاسبات در اختیار یک شبکه اجتماعی خاص قرار می‌گیرد. بنابراین، با تلفیق معادلات محاسبات در مجموعه داده‌های جمع آوری شده در دنیای واقعی، می‌توان تأثیر اجتماعی هر کاربر را اندازه گرفت.

- **طراحی الگوریتم ماکزیمم سازی تأثیر** : الگوریتم ماکزیمم سازی تأثیر [۳۸] برای پیدا کردن kـ موثرترین گره طراحی شده است. اکثر الگوریتم‌های موجود الگوریتم‌های بهبود یافته براساس الگوریتم حریصانه هستند.

- **تجزیه و تحلیل عملکرد در الگوریتم یا مدل مرتبط** : شبیه‌سازی برای اعتبارسنجی عملکرد (به عنوان مثال، دامنه تأثیرگذار، پیچیدگی محاسباتی) الگوریتم‌های پیشنهادی

بر اساس یک مدل انتشار خاص انجام می‌شود. برای یک الگوریتم عالی، این یک دامنه تأثیرگذار بزرگ و پیچیدگی محاسباتی کم است.

جدول ۲.۱: خصوصیات تأثیر

نام	مفهوم
پویا	تأثیر می‌تواند با تجربیات جدید (تعاملات یا مشاهدات) افزایش یا کاهش یابد. همچنین ممکن است در طول زمان از بین برود. در مقایسه با موارد قدیمی، تجربیات جدید از آن زمانی اهمیت بیشتری دارند که ممکن است با پیشرفت زمان ارتباطی نداشته باشند.
پخش شونده	نفوذ، پخش شونده است. به عنوان مثال، A بر B تأثیر می‌گذارد، که ممکن است به نوبه خود بر C نیز تأثیر بگذارد. با این حال، A این را نمی‌داند، C می‌تواند بر اساس میزان اعتماد او به B و میزان اعتماد B به A، به او اعتماد کند. این بدان معنا نیست که اعتماد متعدی است. به دلیل خاصیت پخش شوندگی، می‌توان اطلاعات را از یک عضو به عضو دیگر در یک شبکه اجتماعی منتقل کرد و به همین طریق زنجیره‌های نفوذ بسیاری ایجاد کرد. گسترش نفوذ در شبکه‌های اجتماعی اساس انتشار ’’دهان به دهان’’ اطلاعات برای انسان است.
متعدی	به طور کلی، تأثیر متعدی است. اگر A بر B تأثیر می‌گذارد و B بر Cتأثیر می‌گذارد، این بدان معنی است که A به طور غیرمستقیم بر C تأثیر می‌گذارد.
قابل ترکیب	گسترش نفوذ در طول زنجیره‌های اجتماعی به یک عضو امکان می‌دهد تا روی عضوی که مستقیماً با او در ارتباط نیست هم تأثیر بگذارد. با این حال، هنگامی که چندین زنجیره به طور غیرمستقیم بر یک عضو تأثیر می‌گذارد، آن‌گاه تأثیرگذار باید اطلاعات تأثیرگذاری را بسازد. به عنوان مثال، B از طریق چندین زنجیره در شبکه خود تحت تأثیر A است. در این حالت، A باید اطلاعات نفوذ دریافتی از زنجیره‌های مختلف را تنظیم کند تا تصمیم بگیرد که آیا B بر او تأثیر می‌گذارد یا نه.

نام	مفهوم
قابل‌اندازه‌گیری	سطح نفوذ را می‌توان با یک عدد حقیقی پیوسته اندازه‌گیری کرد، که به آن مقدار نفوذ اشاره دارد. تأثیر نیز می‌تواند نشان‌دهنده عدم قطعیت باشد.
ذهنی	تأثیر ذهنی است. به عنوان مثال، B در مورد یک فیلم نظر می‌دهد. اگر A فکر کند که نظرات B همیشه خوب است، نظر B را می‌پذیرد. با این حال، C ممکن است درباره نظرات B متفاوت فکر کند و ممکن است بررسی را نپذیرد. ماهیت ذهنی تأثیر منجر به شخصی سازی محاسبه تأثیر می‌شود، جایی که تعصبات و سلیقه تأثیرگذار، تأثیر مستقیمی بر ارزش اعتماد محاسبه شده دارد.
نامتقارن	تأثیر به طور معمول نامتقارن است. به عنوان مثال، وقتی می‌گوییم A بر B تأثیر می‌گذارد، به این معنی نیست که B بر A نیز تأثیر می‌گذارد. یک عضو ممکن است بیش از این‌که از عضو دیگری تأثیر بپذیرد، او را تحت تأثیر قرار دهد.
حساس به رویداد	ساختن تأثیر ممکن است مدت‌ها طول بکشد، اما یک رویداد با تأثیر زیاد ممکن است آن را کاملا از بین ببرد. این جنبه از تأثیر در علوم کامپیوتر حتی کمتر مورد توجه قرار گرفته‌است.

کاربردهای کلاسیک:

- **بازاریابی ویروسی**

در بازاریابی ویروسی، برخی معمولاً مجموعه‌ای از افراد با نفوذ را انتخاب می‌کنند و سعی می‌کنند با دادن نمونه آزمایشی یا پرداخت به آن‌ها، ایشان را متقاعد کنند که محصول، خدمات یا نوآوری جدید را قبول کنند [۳۹]. سپس، اجازه می‌دهند آن‌ها محصولات، خدمات یا نوآوری را به دوستانشان در شبکه‌های اجتماعی توصیه کنند. بنابراین، می‌توان از طریق تبلیغ دهان به دهان در شبکه‌های اجتماعی، یک اثر آبشاری ایجاد کرد.

- **یافتن بلاگرهای تأثیرگذار**

وبلاگ نویسان تأثیرگذار نقش مهمی در روند انتشار اطلاعات دارند. آن‌ها توسط مشاغل به عنوان نیروهایی قابل توجه برای ارتقا محصول یا تنزل محصول و توسط رژیم‌های سیاسی ظالمانه به عنوان تهدیدهای جدی برای قدرت آن‌ها شناخته می‌شوند [۴۰].

- **بازیابی اطلاعات**

هر کاربر به عنوان منبع اطلاعاتی در یک شبکه اجتماعی عمل می‌کند، و تأثیر کاربر

برای اعتبار اطلاعات تولید شده معنی‌دار است. در جستجوی افراد وب، تأثیر اجتماعی شاخص اصلی جستجوی تأثیرگذار است [۴۱].

- **تبلیغات آنلاین**

K-بانفوذترین کاربران را انتخاب می‌کند، که می‌تواند به طور خاص کاربرانی را نشان دهد که میل تجاری را گواهی می‌دهند.

- **حداکثر سازی تأثیر**

با راه اندازی یک آبشار انتشار اطلاعات که گره‌های انتخاب‌شده [۴۲] محصول را به دوستان و دوستانِ دوستان خود توصیه می‌کنند. محصول با یک هزینه بازاریابی کوچک به افراد زیادی معرفی می‌شود. [۴۳]

- **مراقبت‌های بهداشتی اجتماعی**

سایت‌های شبکه اجتماعی مراقبت‌های بهداشتی آنلاین [۱۹] به بسترهای محبوب بیماران الکترونیکی برای جستجوی اطلاعات مرتبط با سلامت، به اشتراک گذاشتن درمان، بحث در مورد مسائل پزشکی، درخواست یا ارائه حمایت عاطفی تبدیل می‌شوند.

- **یافتن متخصصان**

تحقیقات در زمینه یافتن خبره [۱۹] با مسئله تحقیق در شناسایی کاربران تأثیرگذار، بسیار مرتبط است، گرچه تعاریف کاربر خبره و تأثیرگذار کاملاً یکسان نیستند.

- **تجزیه و تحلیل احساسات یا استخراج افکار**

تجزیه و تحلیل احساسات [۴۴]، که به عنوان کاوش افکار نیز شناخته می‌شود، از ابزارهای تجزیه و تحلیل متن، پردازش زبان طبیعی و زبان شناسی محاسباتی برای استخراج اطلاعات ذهنی از مجموعه داده‌های داده شده استفاده می‌کند. بر اساس تحلیل تأثیر اجتماعی، می‌توان از تحلیل احساسات برای شناسایی نگرش یک سخنران معین استفاده کرد.

- **شبکه همکاری**

اقدامات تأثیر علمی برای دهه‌ها مورد مطالعه قرار گرفته است. هدف آن نیز کشف ارتباطات روابط نویسندگان است. از دیدگاه اجتماعی، دانشمندان با تأثیر زیاد لزوماً تأثیرگذار نیستند[۴۵].

- **توصیه شخصی**

از آن‌جا که وبلاگ نویسان با نفوذ، اغلب تأثیر بیشتری در تصمیم خرید خوانندگان خود نسبت به تبلیغات دارند، مردم معمولاً به توصیه‌های افراد آگاه و دوستانشان اعتماد بیشتری می‌کنند و براساس نظر آن‌ها عمل می‌کنند [۴۶].

- **انتشار بدافزار**

بر اساس تأثیر اجتماعی هر کدام از رأس شبکه، این استراتژی رئوس بحرانی را در کل شبکه انتخاب می‌کند. سپس، برای تحقق حداکثر اثر مهار کرم با حداقل هزینه، ایمن سازی بر روی رأس انتخاب‌شده اعمال می‌شود [۴۷].

- **تحلیل رفتار**

تأثیر اجتماعی می‌تواند در فعالیت‌های اجتماعی کاربران مانند رفتارهای اجتماعی کاربر منعکس شود.

- **انتقال اطلاعات حساس به زمان**

به خوبی می‌دانیم که نفوذ مبتنی بر علیت در مکانیزم تولید فعالیت، تلویحی و ضمنی است. از این رو یک معیار نفوذی که می‌تواند حجم و سرعت (تاخیر علت) تعامل بین گره‌ها را تصرف کند بسیار مفید خواهد بود زیرا تأثیر نفوذ هدایت شده در OSNها می‌تواند از طریق انتقال اطلاعات حساس به زمان استفاده شود [۴۸].

- **مدیریت داده، مدیریت شبکه و امنیت**

تجزیه و تحلیل تأثیر اجتماعی با شناسایی گره‌هایی که بیشترین betweenness centrality را دارند می‌تواند روی عملکرد شبکه (از طریق تامین منابع)، امنیت (با محدود کردن نظارت بر فعالیت‌های مخرب بالقوه به گروه کوچکی از نامزدها)، و شناسایی ترافیک تأثیر بگذارد[۴۹].

- **یافتن Opinion Leader**

Opinion Leader [۵۰] دارای شخصیت بسیار جذاب توانایی جامع بسیار قوی و موقعیت اجتماعی بالاتری هستند. آن‌ها کسانی هستند که اطلاعات، ایده‌ها و نظرات جدیدی را وارد می‌کنند، سپس آن‌ها را برای توده‌ها پخش می‌کنند.

مدلهای نفوذ

مدل‌های نفوذ معمولاً به عنوان موارد خاص مدل‌های اپیدمی مورد بررسی قرار می‌گیرند. در [۵۱]، نویسندگان مدل‌های اپیدمی را به طور دقیق بررسی کردند. مدل‌های انتشار کلاسیک [۵۲] در شبکه‌های اجتماعی شامل مدل آبشار مستقل (ICM)، مدل آستانه خطی (LTM) و مدل انتشار حرارت (HDM) است.

- مدل آستانه خطی:

در این مدل، یک گره i توسط همسایگان فعال شده‌اش بر اساس یک معیار وزن $w(i,j)$ فعال می‌شود. به منظور اندازه‌گیری احتمال فعال شدن، به طور تصادفی یک آستانه

$\theta_i \in [0,1]$ برای هر گره i انتخاب می‌شود. در صورتی که تعداد همسایگان فعال گره i را با N_i نشان دهیم، رابطه‌ی بین θ_i و $w(i,j)$ به شکل $\sum_{j=1}^{N_i} w(i,j) \geq \theta_i$ خواهد بود. این مدل روش انتشار را به صورت آستانه‌ای توصیف می‌کند. تأثیر از یک فرد به فرد دیگر توسط یک تابع وزن $w(viV)$ ارائه می‌شود. یک فرد فقط زمانی می‌تواند فعال شود که مجموع تأثیراتی که دریافت می‌کند بیش از یک آستانه تصادفی تعیین شده باشد.

- مدل آبشار مستقل:

فرایند انتشار در مراحل گسسته باعث ایجاد یک آبشار فعال‌سازی می‌شود. در این مدل، گره فعال i می‌تواند فقط یکی از همسایگان غیرفعال خود را به دل‌خواه فعال کند. فارغ از این‌که این فعال‌سازی موفق باشد یا خیر، گره i تا پایان فرایند نمی‌تواند برای فعال‌سازی آن گره همسایه، اقدام کند. این مدل شکل انتشار را به روشی احتمالی توصیف می‌کند. یک فرد می‌تواند همسایه خود را با احتمال خاصی تحت تأثیر قرار دهد.

- مدل انتشار حرارت:

انتشار گرما یک دانش رایج در فیزیک است. شباهت زیادی بین انتشار گرما و انتشار اطلاعات در شبکه‌های اجتماعی وجود دارد. در یک شبکه اجتماعی، خالق و پذیرندگان اولیه یک قطعه از اطلاعات به عنوان منابع گرما عمل می‌کنند.

- مدل اپیدمی:

(۴) مدل اپیدمی

مدل‌های اپیدمی به سه دسته طبقه‌بندی می‌شوند که شامل مدل‌های قطعی، تصادفی و مدل‌های مکانی ـ زمانی هستند. مدل‌های اپیدمی قطعی شامل مدل‌های کلاسیک زیر است: مدل SI (مستعد پذیرش‌ـ مسری)، SIS (مستعد پذیرش‌ـ مسری‌ـ مستعد پذیرش) مدل، SIR (مستعد پذیرش‌ـ مسری‌ـ بازیابی) و غیره. مدل‌های اپیدمی تصادفی شامل سه نوع زیر است: مدل زنجیره مارکوف با زمان گسسته (DTMC)، مدل زنجیره مارکوف با زمان پیوسته (CTMC) و مدل معادله دیفرانسیل تصادفی (SDE). مدل‌های اپیدمی زمانی مکانی خودکار سلولی (CA) را برای مدل سازی نفوذ معرفی می‌کنند.

۳.۱ لزوم انجام تحقیق

کاربران رسانه های اجتماعی هر روز چندین ساعت را صرف خواندن، پست گذاشتن وجست‌وجوی اخبار در پلتفرم‌های میکروبلاگینگ می‌کنند رسانه‌های اجتماعی در حال تبدیل شدن به یک ابزار کلیدی برای کشف اخبار می‌باشد.

از دیرباز، تلویزیون، کانال‌های رادیویی و روزنامه‌ها، تنها منابع خبری بوده‌اند. آن‌ها هنوز جزو منابع خبری برتر و قابل اعتماد می‌باشند با این حال روند جدیدتری به سمت منابع

دیجیتال نیز وجود داشته است. نسبت چشم‌گیری از مخاطبان روزنامه‌ها امروزه آن‌ها را به صورت دیجیتال مطالعه می‌کنند و بسیاری از افراد به رسانه‌های اجتماعی به عنوان یک منبع خبری متکی هستند. رسانه‌های اجتماعی این امکان را به شما می‌دهد تا پست‌های آنلاین خود را با یک کلیک ساده منتشر کنید. این امکان موجب شده است تا اخبار جدید ومهم در میکروبلاگ‌ها منتشر شوند. توییتر یکی از محبوب‌ترین پلتفرم‌های میکروبلاگینگ با بیش از ۲۵۰ میلیون کاربر است. قابلیت دسترسی، سرعت و سهولت استفاده موجب شده است تا توییتر ابزار مهمی برای خواندن و اشتراک اطلاعات باشد. با این حال، همین ویژگی موجب شده است تا توییتر و یا هر پلتفرم میکروبلاگینگ به یک منبع مهم تبدیل شود ولی برخی عدم نظارت‌ها موجب شده است تا زمینه‌ای مهم برای انتشار اطلاعات کاذب و نامطلوب در رسانه‌های اجتماعی باشد. بر همین اساس، این می‌تواند منجر به بروز رخدادهای خطرناک و مضری در شرایط حساس شود که موجب بروز اثرات نامطلوب و منفی بر روی افراد و جامعه می‌شود. بسیاری از جست و جو کننده‌های اطلاعات وجود دارند که برای کسب اطلاعات به یک منبع اکتفا نمی‌کنند ولی این یک راه حل خوب نیست زیرا سایر خروجی‌ها یا منابع خبری نیز به رسانه‌های اجتماعی متکی هستند. تلفن‌های هوشمند و توییتر امکان دسترسی هر گونه اطلاعاتی را قبل از دسترسی به تلویزیون به کاربر می‌دهند. با در نظر گرفتن این که رسانه اجتماعی یک گزینه جذاب برای افرادی است که دنبال اخبار جدید هستند، با این حال می‌تواند موجب فریب افراد با انتشار اطلاعات و شایعات کاذب شود[۲۶].

رسانه‌های اجتماعی به عنوان یک پلتفرم برای انجام فعالیت‌های بازاریابی و تبلیغاتی استفاده می‌شوند. سازمان‌ها زمان، پول و منابع زیادی را صرف تبلیغات رسانه‌های اجتماعی کرده‌اند. با این حال، چالش‌هایی در شیوه طراحی تبلیغات رسانه‌های اجتماعی برای جذب موفق مشتریان و انگیزش آن‌ها برای خرید برندهای خود وجود دارد.

رسانه‌های اجتماعی روز به روز جایگاه و اهمیت بیشتری در همه ابعاد زندگی ما پیدا می‌کنند. در واقع، پلتفرم‌های رسانه‌ای اجتماعی یک مکان جدید است که مردم، سازمان‌ها و حتی دولت‌ها می‌توانند از نظر تجاری، اجتماعی، سیاسی و آموزشی با یکدیگر ارتباط برقرار کنند و اطلاعات، افکار، محصولات و خدمات را تبادل کنند. در نتیجه، سازمان‌های سراسر جهان درباره نحوه استفاده از این پلتفرم ها در جذب مشتریان و ایجاد رابطه بازاریابی سودآور تامل کرده‌اند. [۵۳]

برای انجام بسیاری از کارها از جمله تبلیغات و فروش آنلاین، شکل‌دهی و کنترل افکار عمومی، تأثیرگذاری در جریان‌های سیاسی و اجتماعی، کنترل شایعات و بسیاری دیگر از این نوع موارد، نیاز به شناسایی افراد تأثیرگذار داریم.

۴.۱ اهداف تحقیق

هدف این تحقیق یافتن افراد تأثیرگذار در شبکه اجتماعی توییتر با ارائه‌ی تعریفی مناسب برای مفهوم تأثیر و همچنین انتخاب مناسب‌ترین مؤلفه‌ها برای اندازه‌گیری میزان تأثیر می‌باشد. توضیحات کامل‌تری در این مورد در فصل ۳ آمده است.

۵.۱ پیش‌فرض‌های مسئله

پیش‌فرض‌هایی که برای این سیستم در نظر گرفته شده است، عبارتند از:

- توییت‌ها همگی به زبان انگلیسی باشند.
- از هر کاربر موجود در گراف حداقل یک توییت موجود باشد

۶.۱ پایگاه داده

پایگاه داده‌های استفاده شده در این تحقیق، archiveteam twitter stream[۵۴، ۵۵] برای توییت‌ها و Twitter Social Graph [۵۶] برای گراف کاربران در شبکه است. پایگاه داده توییت‌ها حاصل جمع‌آوری تمام توییت‌های دو ماه ابتدایی سال ۲۰۱۲ در لحظه انتشار هستند که شامل مشخصات کامل توییت‌ها در همان لحظه هستند. پایگاه داده کاربران هم شامل گراف کامل رابطه‌ی دنبال‌کنندگان کاربران در سال ۲۰۱۰ شبکه اجتماعی توییتر می‌باشد. توضیحات کامل‌تری در رابطه با پایگاه داده در فصل ۳ آمده است.

۷.۱ جمع‌بندی و ساختار پایان‌نامه

در این فصل شبکه‌های اجتماعی را از لحاظ مختلف بررسی کردیم، هدف و کاربرد این شبکه‌ها بیان شده و به موضوع تحلیل شبکه‌های اجتماعی و تحلیل تأثیر اجتماعی پرداختیم. در فصل دوم کارهای مرتبط را از لحاظ مختلف مورد بررسی قرار می‌دهیم و سپس به بیان روش پیشنهادی در فصل سوم می‌پردازیم. پس از آن، در فصل چهارم پیاده‌سازی روش پیشنهادی و نتایج آن را بررسی می‌کنیم و در انتها به نتیجه‌گیری در فصل پنجم می‌پردازیم.

فصل ۲

مروری بر کارهای پیشین

۱.۲ مقدمه

در این بخش به بررسی کارهای مرتبط انجام‌شده خواهیم پرداخت. ابتدا آن‌ها را از لحاظ معیارهای ارزیابی مورد استفاده بررسی می‌کنیم، سپس مدل‌های ارزیابی آن‌ها را مقایسه کرده و در نهایت آن‌ها را بر اساس الگوریتم‌های حداکثر سازی تأثیر ارزیابی خواهیم‌کرد.

۲.۲ معیارهای ارزیابی تأثیر اجتماعی

معیارهای ارزیابی موجود برای تأثیر اجتماعی شامل معیارهای مرکزیت، معیارهای رتبه بندی توپولوژیک پیوند، اندازه‌گیری آنتروپی و غیره است.

۱.۲.۲ معیارهای مرکزیت

مرکزیت، یک مفهوم مهم در مطالعه شبکه‌های اجتماعی است. از نظر مفهومی، مرکزیت نحوه قرارگیری یک فرد در یک شبکه اجتماعی را اندازه‌گیری می‌کند. ابزارهایی که معمولا در این زمینه مورد استفاده قرار می‌گیرند نظریه گراف و تحلیل شبکه هستند. تاکنون معیارهای مرکزیت متنوعی ارائه شده است که به طور گسترده در تجزیه و تحلیل تأثیرات اجتماعی

استفاده می‌شود، از جمله degree-centrality، closeness-centrality، betweenness-centrality، eigenvector-centrality و Katz-centrality [۵۷].

۲.۲.۲ معیارهای رتبه‌بندی پیوند توپولوژیک

به جز معیار مرکزیت بردار ویژه، اکثر معیارهای مرکزیت ذکر شده تغییر گره‌ها را نادیده می‌گیرند، یعنی آن‌ها معتقدند همه گره‌ها به طور مساوی در اندازه‌گیری‌ها سهیم هستند. با این حال، انواع گره‌ها نقش مهمی در شبکه‌های اجتماعی دارند.

اتصال به یک گره با مرکزیت بالا از اتصال به گره با همسایه‌های کم، ارزش بیشتری دارد. موتورهای جستجوی وب از طریق استفاده از الگوریتم‌های جستجوی موضوع توسط Hyperlink (HITS) [۵۸] و PageRank [۵۹] از رتبه بندی توپولوژیکی پیوند استفاده می‌کنند.

۳.۲.۲ معیارهای آنتروپی

از آن‌جا که آنتروپی یک ابزار موثر برای توصیف پیچیدگی و عدم قطعیت نفوذ اجتماعی است، در شبکه‌های اجتماعی به طور گسترده‌ای مورد استفاده قرار گرفته‌است. پنگ و همکاران [۶۰] دو مفهوم، آنتروپی دوستان و آنتروپی فرکانس تعامل را اختراع کردند تا تأثیر اجتماعی در شبکه‌های اجتماعی موبایل را اندازه‌گیری کنند. سان و نگ [۶۱] از آنتروپی گراف برای اندازه‌گیری نفوذ اجتماعی در شبکه اجتماعی استفاده کردند. Sathanur و Jandhyala [۴۸]، آنتروپی انتقال را معرفی کردند تا تأثیر مبتنی بر علیت را اندازه‌گیری کنند. استیگ و گالستیان [۶۲] نیز تأثیر اجتماعی در شبکه‌های اجتماعی را با استفاده از آنتروپی انتقال اندازه‌گیری کردند. هی و سایرین [۶۳] بودند که آنتروپی انتقال را در شبکه‌های اجتماعی آنلاین معرفی کردند.

۴.۲.۲ معیارهای دیگر

چن و همکاران [۶۴] برای اندازه‌گیری تأثیر اجتماعی در شبکه اجتماعی از چهار معیار (به عنوان مثال، تعداد دنبال‌کنندگان، کیفیت دنبال‌کنندگان، کیفیت توییت‌ها و شباهت مورد علاقه) استفاده کرد. لی و ژیلت [۴۵] از سه معیار (به عنوان مثال، حداکثر تعداد خوانندگان در هر مقاله، تعداد کل خوانندگان و R-Index) برای اندازه‌گیری تأثیر دانشمندان در سیستم‌های رسانه‌های اجتماعی دانشگاهی استفاده کردند. هوی و گریگوری[۶۵] از چهار معیار (به عنوان مثال، تعداد دنبال‌کنندگان، میزان مرتبط بودن، نظرات و دنبال‌کنندگان) برای اندازه‌گیری تأثیر در فضای وبلاگ استفاده کردند. از سوی دیگر آگاروال و همکاران [۱۶] نیز از چهار معیار مانند شناخت، تولید فعالیت، تازگی و شیوایی برای سنجش تأثیر بلاگرها استفاده کرده‌اند. چا و همکاران [۱۵] برای اندازه‌گیری تأثیر در توییتر از سه معیار مانند درجه ورودی

(دنبال‌کنندگان، in-degree، followers)، بازتوییت (retweet) و ذکر نام (mention) استفاده کرد. همچنین پنگ و همکاران [۳۸، ۶۶] از عوامل مختلفی برای اندازه‌گیری تأثیر اجتماعی در شبکه‌های اجتماعی تلفن همراه استفاده کرده‌اند، مانند تعداد گره‌ها، تعداد فعل و انفعالات بین گره‌ها و فعالیت گره‌ها.

۳.۲ مدل‌هایی برای سنجش میزان تأثیر اجتماعی

اخیرا تلاش‌هایی برای مدل‌سازی تأثیر اجتماعی در شبکه‌های اجتماعی به کار گرفته شده است. سه نوع اصلی به شرح زیر هستند [۶۷]. (۱) مبتنی بر محتوا: برخی از طرح‌ها [۱۵، ۶۸، ۶۹] در ارزیابی تأثیر اجتماعی موضوع و محتوا محور بوده‌اند. در این مدل‌ها، تأثیر اجتماعی با شمارش میزان انتشار اطلاعات مربوط به یک موضوع در شبکه‌ها اندازه‌گیری شد[۷۰]. (۲) مبتنی بر روابط: برخی از طرح‌ها [۱۴، ۱۹، ۷۱] بر روی تأثیر اجتماعی توجهی به موضوع و یا محتوا نداشتند. در این مدل‌ها تأثیر اجتماعی از طریق قدرت نسبی افراد در شبکه‌های اجتماعی و روابط آن‌ها در شبکه‌های اجتماعی اندازه‌گیری شد. (۳) مبتنی بر محتوا و روابط: برخی از طرح‌ها [۷۲، ۷۳، ۷۴] وابسته به هر دو مدل مبتنی بر محتوا و مبتنی بر روابط اجتماعی و تعاملات کاربران برای ارزیابی تأثیر اجتماعی هستند.

۱.۳.۲ مبتنی بر محتوا

با توجه به نوع مدل‌های موجود، بسیاری از محققان بر بازاریابی ویروسی تمرکز کرده‌اند، این موارد به شرح زیر ذکر شده است.

Cha و همکاران، روشی را برای ارزیابی تأثیر کاربر در توییتر ارائه کردند. آن‌ها برای اندازه‌گیری تأثیر کاربر از سه معیار تعداد دنبال‌کنندگان، تعداد نام بردن‌ها [۱] و تعداد ریتوییت‌ها[۲] استفاده کردند. در همین زمینه دینگ و همکاران [۶۸] یک مدل مولد احتمالی آگاه از زمان (TPGM) را برای ارزیابی قدرت نفوذ با در نظر گرفتن بازه‌ی زمانی، رابطه پیروی و محتوای پست ایجاد کرده‌اند.

حاجیان و وایت [۷۵] هم مدلی را برای سنجش نفوذ در یک شبکه اجتماعی ارائه دادند که هم بر اساس ساختار شبکه و هم بر تعامل بین گره‌های شبکه بنا شده است. آن‌ها شاخصی به نام اندازه‌ی نفوذ (MOI[۳]) را معرفی کردند که براساس اندازه تأثیری که پست‌های کاربر بر همسایگی وی می‌گذارد، اندازه‌گیری می‌شود.

بی و همکاران [۷۶] مدل مخلوطی از بیزین برنولی و چندجمله‌ای، به نام FLDA، را ارائه نمودند که کشف موضوع محتوا و همچنین تحلیل نفوذ اجتماعی را در یک فرآیند مولد ترکیب

[1] Mention
[2] retweet
[3] Magnitude of Influence

می‌کند. هم‌چنین سان و نگ [۶۱]یک مدل گراف را برای نشان دادن روابط بین پست‌های آنلاین یک موضوع ارائه دادند. آن‌ها از سه روش برای اندازه‌گیری تأثیر پست‌های آنلاین با متمایز کردن شروع کننده‌ها (یک شروع کننده توسط بسیاری دیگر دنبال می‌شود، بنابراین باید تأثیر خاصی داشته باشد) و اتصال دهنده‌ها (اتصال دهنده همچنین هنگامی که شروع کننده‌ها را به یکدیگر متصل می‌کند نیز تأثیرگذار است) استفاده کردند.

علاوه بر این، بسیاری از محققان به کاربردهای شبکه‌های اجتماعی علاقه‌مند بودند (به عنوان مثال، شبکه‌های استنادی و همکاری) که مدل‌های مرتبط به شرح زیر هستند:

تانگ و همکاران روش انتشار قرابت موضوعی (TAP) [۳۰] را برای مدل سازی تأثیر اجتماعی در سطح موضوع ارائه دادند. نویسندگان، یک الگوریتم یادگیری موازی بر اساس مدل MapReduce برای روش خود ایجاد کردند. گو و همکاران [۲۵] برای سنجش قدرت نفوذ در شبکه‌های استنادی، یک مدل مولد به نام مدل موضوعی فرآیند برنولی (BPT) را ارائه دادند. آن‌ها این مجموعه را در دو سطح مدل‌سازی کردند: سطح سند و سطح استناد. به تقلید از این کار، دیتز و همکاران [۲۴] مدل نفوذ استنادها را برای مطالعه تأثیر نقل قول‌ها در مجموعه‌ای از نشریات ایجاد کرد. مدل کپی برداری با افزودن حداقل تعداد متغیرهای تصادفی، مدل LDA را تعمیم داد. گریش و بلی [۷۷] برای سنجش تأثیر یک مقاله علمی، مدل تأثیر سند (به اختصار DIM) را ارائه دادند که از زنجیره‌ی مارکف برای توضیح توزیع استفاده می‌کند.

در ادامه دستاوردهای موجود تانگ و همکاران [۷۸] یک مدل گراف آگاه به انطباق، به نام تلاقی، برای تفکیک تأثیرات انواع مختلف انطباق‌ها، مانند انطباق فردی، مطابقت نظیر و انطباق گروهی، ارائه داد. هرتسیگ و همکاران [۷۹] نیز مدل تأثیر نویسنده‌ـخواننده (ARI) را برای سنجش تأثیر کاربران مختلف بر دیگران با استفاده از تحلیل گذشته نگر از دیدگاه یک خواننده معمولی ارائه دادند. یک گراف استنادی مبتنی بر محتوا برای تحقق بخشیدن به مدل ساخته شد. به همین طریق ژانگ و همکاران [۸۰] تأثیر انطباق را با استفاده از یک تابع مطلوب مبتنی بر تئوری انطباق از روانشناسی اجتماعی رسمیت بخشید، و سپس یک مدل گرافیکی احتمالی، که به عنوان مدل انطباق نقش (RCM) نامیده می‌شود، برای مدل سازی تأثیر انطباق آگاهانه بین کاربران با استفاده از تابع ابزار ارائه دادند.

به غیر از مدل‌های فوق، بسیاری از محققان نیز بر دامنه‌های کاربردی زیر تمرکز کرده‌اند، مانند جستجوی جمعی [۲۳]، بازیابی اطلاعات [۸۱]، یافتن وبلاگ نویسان با نفوذ [۱۶]، تجزیه و تحلیل احساسات [۴۴، ۶۵]، مدیریت داده [۴۹]، جستجوی موضوع تأثیرگذار شخصی [۸۲] و غیره.

سانگ و خو [۲۳] نیز برای حل مسئله استخراج تأثیرگذار مبتنی بر محتوا در وب‌سایت‌های اشتراک اجتماعی، یک مدل احتمالی چند حالته ارائه دادند. همچنین کوی و همکاران [۸۱] رویکرد فاکتورسازی ماتریس غیر منفی فاکتور ترکیبی (HF-NMF) را برای مدل سازی تأثیر

اجتماعی در سطح مورد، ارائه دادند. آن‌ها پیشوندها را با استفاده از ماتریس شباهت کاربر_کاربر و ماتریس توزیع موضوع برای پیش بینی تأثیر اجتماعی، از کاربران و پست ها ساختند و یک روش شیب پیش‌بینی شده برای حل مسئله HF-NMF ابداع کردند. آگاروال و همکاران [۱۶] مدلی برای تعیین کمیت تأثیر Blog-post در یک جامعه ارائه داد. در این مدل، آن‌ها از چهار عامل، از جمله شناخت، تولید فعالیت، تازگی و سخنوری، برای اندازه‌گیری تأثیر هر بلاگ_پست استفاده کردند.

هوی و همکاران [۶۵] روشی برای محاسبه تأثیر در فضای وبلاگ ارائه دادند. آن‌ها برای اندازه‌گیری تأثیر هر وبلاگ از چهار معیار، تعداد دنبال‌کنندگان، میزان مرتبط بودن، نظرات و دنبال‌کنندگان استفاده کردند. این جستجو باعث گردید تا کورتلیس و همکاران [۴۹] برای ارزیابی مرکزیت میانی یک گره، شاخص مرکزیت گراف به نام مرکزیت k_مسیر را معرفی کنند و یک الگوریتم k مسیر تصادفی تقریبی برای تخمین مقدار آن برای همه گره‌ها ارائه دهند.

لی و همکاران [۴۴] برای اندازه‌گیری تأثیر عقاید اجتماعی، مدل تأثیر نظر در سطح موضوع (TOIM) را ارائه نمودند. آن‌ها فاکتور موضوع، نظرات کاربران و تأثیر اجتماعی را در یک مدل احتمالی واحد با دو مرحله فرآیند یادگیری گنجاندند و برای تعیین کمیت تأثیر نظر بین دو کاربر، شاخص توافق نظر متریک را ارائه دادند. آن‌ها همچنین در پژوهشی دیگر [۸۲] با استفاده از شناسایی روابط بین گره‌های موضوع و گره‌های نماینده، شاخص کاربری مربوط به موضوع را به نماینده برای ایجاد تأثیر محلی حساس به موضوع در جوامع کوچک در شبکه‌های اجتماعی ایجاد کردند.

۲.۳.۲ مبتنی بر روابط

به منظور برآورد قدرت رابطه، شیانگ و همکاران [۷۱] مدلی بدون نظارت مبتنی بر فعالیت تعامل و شباهت کاربر ایجاد کردند. آن‌ها یک مدل متغیر پنهان مبتنی بر لینک (به اختصار LVM) فرموله کردند و یک روش بهینه سازی صعود مختصات را برای استنتاج طراحی کردند. با این حال، تجزیه و تحلیل قدرت تأثیر میکروسکوپی در کار آن‌ها نادیده گرفته می‌شود.

دومینگوس و ریچاردسون [۱۴] نیز تأثیر اجتماعی در شبکه‌های مشتریان را بررسی کردند و مدلی را برای شناسایی تأثیر مشتری‌ها بین یکدیگر، ارائه دادند. آن‌ها همچنین یک مدل احتمالی ساختند تا گسترش نفوذ را استخراج کنند. تانگ و همکاران [۱۹] یک روش ترکیبی برای اندازه‌گیری تأثیر کاربر از طریق ترکیب رویکردهای مبتنی بر محتوا و مبتنی بر شبکه در یک جامعه آنلاین، پیشنهاد کردند.

سونگ و همکاران [۵۰] یک الگوریتم رتبه بندی، به نام InfluenceRank-1 ، برای شناسایی رهبرانی که نظرات جدیدی دارند و همچنین در شبکه بسیار تأثیرگذار هستند، پیشنهاد کردند. در الگوریتم پیشنهادی، آن‌ها از اندازه شباهت کسینوسی برای اندازه‌گیری جدید بودن اطلاعات ورودی‌ها یا وبلاگ‌ها استفاده کردند. سپس InfluenceRank برای شناسایی رهبران نظر

(opinion-leaders) با در نظر گرفتن اهمیت و تازگی اطلاعات گره‌ها در شبکه پیشنهاد شد. دسته دیگری از محققین چون Kayes و همکاران [۴۰] برای اندازه‌گیری نمرات تأثیر نسبی وبلاگ نویسان در شبکه، روش تجمیع مرکزیت (به اختصار CAM) را پیشنهاد کردند. آن‌ها برای اندازه‌گیری تأثیر اجتماعی وبلاگ نویسان از شش معیار مرکزیت شبکه مانند degree، betweenness، communicability، closeness، eigenvector و hub استفاده کردند. وانگ و همکاران [۸۳] مدلی را پیشنهاد کردند، به نام مدل تأثیر اجتماعی پویا (به اختصار DSIM)، این رویکرد چنین فرایندهای تأثیرگذار اجتماعی را شبیه‌سازی می‌کند که افراد هنگام برقراری ارتباط و تبادل نظر با دیگران، نگرش خود را به طور پویا تغییر می‌دهند.

استیگ و گالستیان [۶۲] تأثیر اجتماعی در شبکه‌های اجتماعی را با استفاده از آنتروپی انتقال اندازه‌گیری کردند. آن‌ها محتوای تولید شده توسط کاربر را در فضایی با ابعاد بالا نشان می‌دهند به طوری که دنباله‌ای از پست‌های ارسال‌شده توسط کاربران به یک سری زمانی در این فضا نگاشت می‌شود، و معیارهای انتقال محتوا را به آن سری‌های زمانی اعمال می‌کنند تا تأثیر مستقیم را در بین کاربران کشف و تعیین کنند.

لی و ژیلت [۴۵] تأثیر دانشمندان را درپلتفرم‌های اجتماعی دانشگاهی بررسی کردند. آن‌ها تأثیر علمی دانشمندان را بر اساس تأثیر علمی انتشاراتشان با استفاده از سه معیار مختلف (یعنی تعداد کل خوانندگان، حداکثر تعداد خوانندگان در هر مقاله، شاخص R) ارزیابی کردند و تأثیر اجتماعی آن‌ها را با استفاده از معیارهای مرکزیت شبکه بررسی کردند.

۳.۳.۲ مبتنی بر محتوا و روابط

با توجه به نوع مدل‌های موجود، بسیاری از محققان بر روی «یافتن متخصص» (expert finding) متمرکز شده‌اند، این موارد به شرح زیر ذکر شده است.

پال و همکاران [۷۲] روشی را برای شناسایی مراجع موضوعی در میکروبلاگ ها پیشنهاد کردند. برای کمی کردن میزان تأثیر مبتنی بر محتوا و روابط و شناسایی مراجع موضوعی در توییتر، از اطلاعات مربوط به ریتوییت حساس به موضوع استفاده شد. با این حال، این روش برای سایر شبکه‌های اجتماعی مناسب نبود. وانگ و همکاران [۲۱] مدل ارزیابی دقیق تأثیر مبتنی بر ویژگی (FBI) در شبکه‌های اجتماعی آنلاین را ارائه دادند. آن‌ها با کاوش در دو عامل اساسی (امکان تأثیرگذاری بر دیگران و اهمیت خود کاربر)، تأثیر اجتماعی اولیه کاربر را ایجاد کردند.

لیو و همکاران [۸۴] با در هم آمیختن اطلاعات متنی و همچنین اطلاعات پیوند در شبکه‌های ناهمگن، یک مدل مولد احتمالی برای استخراج تأثیر سطح موضوع بین گره‌ها ارائه داد.

علاوه بر این، بسیاری از محققان به کاربردهای شبکه‌های اجتماعی علاقه‌مند بوده‌اند (به عنوان مثال، بازاریابی ویروسی و یافتن وبلاگ نویسان با نفوذ). مدل‌های مربوط به شرح زیر

ذکر شده‌اند.

Goyal و همکاران [۷۴] برای تعیین کمیت قدرت تأثیر با استفاده از احتمال تأثیر، دو مدل استاتیک و وابسته به زمان را پیشنهاد دادند. آن‌ها برای یادگیری این احتمالات از مدل آبشار مستقل استفاده کردند، به این شکل که هرچه کاربر v موارد بیشتری (پست و غیره) از کاربر u دریافت کند، احتمال تأثیر کاربر u بر کاربر v افزایش می‌یابد. شهود این است که هرچه تعداد بیشتری از دوستان کاربر عملی را انجام دهند، احتمال انجام آن کار توسط این کاربر نیز بیشتر می‌شود.

He و دیگران [۶۳] روشی را برای شناسایی تأثیر همتا در شبکه‌های اجتماعی آنلاین ارائه دادند. در این روش، با تعیین اطلاعات جهت دار یا جریان اطلاعات متقابل مشروط بین دو منبع سیگنال با استفاده از آمار سیگنال غیر پارامتری چند متغیره، آنتروپی انتقال برای شناسایی تأثیر همتا در شبکه‌های اجتماعی آنلاین معرفی شد.

در ادامه این دستاوردها آرال و واکر [۸۵] روشی را ارائه دادند که از آزمایشات تصادفی در vivo برای شناسایی تأثیر و حساسیت در شبکه‌ها استفاده می‌کند. در این روش، آن‌ها تأثیر روابط دوگانه را با استفاده از مدل خطرات متناسب با یک شکست مداوم زیر تخمین زدند. پنگ و همکاران [۶۶] دو عامل را برای ارزیابی تأثیر هر گره معرفی کردند. یک عامل درجه صمیمیت است که برای انعکاس نزدیکی بین کاربران استفاده شده است. عامل دیگر درجه فعالیت بود که برای تعیین میزان فعال بودن گره مورد استفاده قرار گرفت. Tang و همکاران با استفاده از اطلاعات برچسب ارائه شده توسط یک شبکه اجتماعی آنلاین، [۸۶] اندازه‌گیری تأثیر مبتنی بر محتوا و روابط را پیشنهاد دادند. Bakshy و همکاران [۱۷] مقدار نفوذ مبتنی بر محتوا و روابط را به صورت نمره‌ی نفوذ در نظر گرفتند و همچنین با جمع آوری تمام نمرات تأثیرگذاری کاربر، تأثیر کلی او را درنظر گرفته و برای پیش‌بینی تأثیر کلی کاربر، مدل درخت نفوذ جداگانه‌ای را با برخی ویژگی‌ها ساختند.

دنگ و دای [۸۷] یک مدل تخمین احتمال retweet را برای تخمین تأثیر دو به دو در توییتر بر اساس تئوری بیزی ارائه دادند. آن‌ها تأثیر دو به دو را به عنوان احتمال retweet تعریف کردند، با مشاهده داده‌های تاریخچه retweet، تأثیر را تخمین زدند و یک قاعده بیزی را برای ارزیابی احتمال retweet معرفی کردند. وانگ و همکاران [۸۸] نیز یک مدل گراف دو به دو را برای مدل سازی تأثیر اجتماعی در شبکه‌های اجتماعی ارائه دادند. آن‌ها از یک احتمال حاشیه‌ای برای تعریف تأثیر دو به دو بین کاربران استفاده کردند. Saez-Trumper و همکاران [۸۹] هم روشی را برای یافتن الگوهای روند در شبکه‌های اطلاعاتی با توجه به موضوع خاص مورد علاقه ارائه نمودند. آن‌ها یک استراتژی رتبه‌بندی را پیشنهاد کردند که بر توانایی برخی کاربران برای انتشار ایده‌های جدیدی که در آینده موفق خواهند بود، متمرکز است و ویژگی‌های زمانی گره‌ها و لبه‌های شبکه را با الگوریتم مبتنی بر Pagerank ترکیب کردند تا طراحان بتوانند یک موضوع خاص را پیدا کنند.

به غیر از مدل‌های فوق، بسیاری از محققان نیز بر دامنه‌های کاربردی زیر متمرکز شده‌اند،

مانند تجزیه و تحلیل رفتار [۷۳، ۹۰]، بازیابی اطلاعات [۴۱]، توصیه [۱۸]، انتقال اطلاعات حساس به زمان [۴۸]، مراقبت‌های بهداشتی آنلاین [۲۰]، داده کاوی [۹۱] و غیره.

از دیگر یافته‌های محققین می‌توان به پان و همکاران [۹۰] اشاره کرد که یک مدل تأثیر پویا ارائه دادند، آن‌ها نشان دادند که چگونه می‌توان این موارد را بر روی انواع سیگنال‌های اجتماعی اعمال کرد تا نتیجه بگیرد که چگونه موجودیت‌ها در شبکه‌ها بر یکدیگر تأثیر می‌گذارند. کواک و همکاران [۷۳] نیز از سه معیار مختلف تأثیر مانند PageRank، تعداد دنبال‌کنندگان و تعداد ریتوییت‌ها برای محاسبه تأثیر دو به دو استفاده کردند و از الگوریتم PageRank برای یافتن کاربران تأثیرگذار بهره گرفتند. با این حال، معیارهای مدل برای درک تفاوت بین دوستان بسیار ساده هستند. جابور و همکاران [۴۱] روشی را برای سنجش تأثیر اجتماعی بر اساس مدل شبکه بیزی، با استفاده از برخی عوامل مرتبط، مانند اهمیت اجتماعی کاربران فعال میکروبلاگ‌ها و اندازه زمانی توییت‌ها، پیشنهاد کردند.

از سوی دیگر هوانگ و همکاران نیز بسته به تعداد دنبال‌کنندگان و حساسیت آن‌ها در کشف موارد خوب،[۱۸] تأثیر اجتماعی افراد را به طور کمی بررسی کردند. همچنین Sathanur و Jandhyala [۴۸] مدلی را برای سنجش تأثیر مبتنی بر علیت جهت دار با معرفی آنتروپی انتقال ارائه دادند. آن‌ها برای توصیف ماهیت نفوذ، گاه به گاه بین دو کاربر در شبکه‌های اجتماعی آنلاین از مشخصات آنتروپی انتقال تاخیر جابجایی (DTE) مورد استفاده قرار دادند.

۴.۲ الگوریتم حداکثر سازی تأثیر

در تدوام این تحقیقات دومینگوس و ریچاردسون [۱۴] اولین کسانی بودند که حداکثر تأثیر را به عنوان یک روش برای بازاریابی ویروسی پیشنهاد دادند. کمپ و همکاران [۳۴] نیز اولین کسانی هستند که به حداکثر رساندن تأثیر را به عنوان یک مسئله بهینه سازی گسسته رسمیت می‌دهند و اثبات می‌کنند که این مسئله NP-hard است. مطابق ادبیات مرتبط، مسئله حداکثر تأثیر [۹۲، ۹۳] به صورت زیر فرموله شده است: اگر یک شبکه اجتماعی که به عنوان گراف $G = (VE)$ و یک عدد غیر منفی k داشته باشیم، به شکلی که گره‌ها کاربرهای گراف بوده و یال‌ها احتمال تأثیرگذاری در بین کاربران، مسئله حداکثرسازی تأثیر یافتن مجموعه‌ای از k گره است که مقیاس انتشار نفوذ را در شبکه اجتماعی تحت یک مدل انتشار مشخص به حداکثر می‌رساند.

در این بخش، ما انواع الگوریتم حداکثر سازی تأثیر را شرح داده و سپس الگوریتم‌های موجود را مرور می‌کنیم.

انواع الگوریتم حداکثر سازی تأثیر بطور کلی به به شکل زیر هستند:

- **روش‌های حریصانه**

الگوریتم حریصانه، الگوریتمی است که در هر مرحله بهترین انتخاب محلی را برمی‌گزیند

و امیدوار است که با همین روش به بهترین جواب ممکن دست‌یابد. در بسیاری از موارد، یک استراتژی حریص به طور کلی به یک راه حل مطلوب نمی‌رسد، اما ممکن است راه حل‌های محلی بهینه که تقریبی از یک راه حل کلی مطلوب است را در یک زمان معقول ارائه دهد.

- **روش‌های اکتشافی**

در علوم کامپیوتر، هوش مصنوعی و بهینه سازی ریاضی، یک الگوریتم ابتکاری [۹۴] برای حل سریعتر مسئله در صورت کند بودن روش‌های کلاسیک یا یافتن راه‌حل تقریبی در صورت عدم موفقیت روش‌های کلاسیک طراحی شده است. این امر از طریق بهینه بودن، کامل بودن، دقت یا درستی در معاملات حاصل می‌شود. به نوعی می‌توان آن را میانبر دانست.

- **روش‌های دیگر**

علاوه بر دو نوع عمده، بسیاری از محققان الگوریتم‌های مبتنی بر رأی گیری و الگوریتم‌های ترکیبی را برای بهبود عملکرد و مقیاس پذیری روش‌های موجود حریصانه و اکتشافی ارائه داده‌اند.

۱.۴.۲ روش‌های حریصانه

در این زمینه وانگ و فنگ [۹۵] بر اساس استراتژی انتخاب گره مبتنی بر پتانسیل، یک الگوریتم حریصانه را پیشنهاد کردند. این الگوریتم ممکن است در فاز اولیه مطلوب نباشد، اما می‌تواند گره‌های بیشتری را در مرحله بعدی انتشار، بیابد. همچنین لی و همکاران [۹۶] نیز یک الگوریتم حریصانه آگاه به انطباق، به نام CINEMA، برای حل مسئله حداکثر تأثیر ارائه داده‌اند. این الگوریتم برپایه‌ی یک مدل آبشاری آگاه به انطباق جدید ساخته شده است. وانگ و همکاران [۹۷] یک الگوریتم حریصانه مبتنی بر انجمن برای مسئله حداکثر تأثیر پیشنهاد دادند. به منظور کاهش هزینه از نظر زمان اجرا، آن‌ها ابتدا انجمن‌ها را بر اساس مدل IC شناسایی کردند. در مرحله دوم آن‌ها K ـ بالاترین گره‌ها را در انجمن‌ها با استفاده از تکنیک‌های استخراج(mining) به دست آوردند. آن‌ها تابع هزینه را به منظور بهینه‌سازی انتساب جامعه در شبکه‌های سیار را نیز فرموله کردند.

در راستای این دستاوردها ژانگ و همکاران [۹۸] یک الگوریتم حریصانه، برآورد ـ انتظار حریصانه (GEE)، بر اساس تقریب به انتظار ارائه نمودند. لیو و همکاران [۹۹] نیز با استفاده از قابلیت پردازش موازی واحد پردازش گرافیک (GPU) چارچوبی را برای افزایش سرعت حداکثر سازی تأثیرگذاری ارائه دادند. در ابتدا برای جلوگیری از محاسبه زائد، گراف اجتماعی را به گراف جهت‌دار بدون دور (DAG) تبدیل کردند. سپس یک الگوریتم پیمایش پایین به بالا (BUTA) به کار گرفته شد و با مدل برنامه نویسی CUDA به GPU ترسیم گردید. کمپ و

همکاران [۳۴] ثابت کردند که مشکل به حداکثر رساندن تأثیر، NP سخت است و یک الگوریتم حریص صعود کننده را بر اساس مدل آستانه خطی و مدل آبشار مستقل پیشنهاد دادند و تقریب $(\frac{1}{e} - 1)$ از راه حل بهینه را ارائه کردند. Ma و همکاران [۱۰۰] نیز یک الگوریتم حریصانه k-step برای انتخاب نامزدهای بازاریابی با استفاده از یک فرآیند انتشار گرما در یک گراف جهت‌دار و یک گراف بدون جهت پیشنهاد کردند. به همین قیاس ژو و همکاران [۱۰۱] هم یک الگوریتم حریصانه بر اساس ترجیحات کاربران برای استخراج تأثیرگذارترین گره‌های شبکه با موضوع مشخص طراحی کردند. این الگوریتم در دو مرحله کار می‌کند، در مرحله اول، یک مدل فضای برداری برای محاسبه ترجیحات کاربر طراحی شده است. در مرحله دوم، الگوریتمی حریصانه برای یافتن تأثیرگذارترین گره‌ها اتخاذ می‌شود.

Estevez و همکاران [۱۰۲] الگوریتمی حریصانه را برای حل مسئله حداکثر سازی تأثیر اجتماعی ارائه دادند. این الگوریتم گره‌ای را با بالاترین ''درجه‌های کشف نشده'' انتخاب می‌کند، به محض انتخاب یک گره، گره و همسایگان آن به عنوان ''پوشانده شده'' برچسب گذاری می‌شوند. این روش تا انتخاب گره‌های k ادامه می‌یابد. Goyal و همکاران [۱۰۳] با در اختیار داشتن چندین بهینه سازی هوشمندانه، الگوریتم SIMPATH را برای همین مسئله تحت مدل آستانه خطی پیشنهاد داد. این روش گسترش نفوذ را با کاوش در مسیرهای ساده در همسایگی محاسبه می‌کند. لی و چونگ [۱۰۴] برای به حداکثر رساندن تأثیر، روش حریصانه‌ای را براساس گسترش نفوذ ۲ مرحله‌ای (GIS) پیشنهاد کردند. Ma و Ma [۱۰۵] یک الگوریتم دو مرحله‌ای به نام حریصانه مبتنی بر نامزدها (CBG) را پیشنهاد کردند، که در ابتدا n گره با بیشترین تأثیرگذاری را به عنوان نامزدها با روش‌های ابتکاری انتخاب می‌کند، و سپس بذرها را با الگوریتم حریصانه از نامزدها انتخاب می‌کند. ژانگ و همکاران [۱۰۶] یک الگوریتم حریصانه قابل توسعه برای یافتن مجموعه‌ای از کاربران با حداقل کارایی برای تأثیر بخشی خاص از کاربران در شبکه‌های پیچیده ارائه داد. اوک و دیگران [۱۰۷] یک چارچوب الگوریتمی، PaS[1] را طراحی کردند که PaS از دو مرحله تشکیل شده است: (الف) تقسیم گراف به چند خوشه و (ب) بذر در هر خوشه. سپس، آن‌ها الگوریتمی حریصانه به نام Pas کاربردی (PrPaS) را برای دستیابی به تخصیص بودجه مورد نظر در مرحله بذر ارائه دادند.

۲.۴.۲ روش‌های اکتشافی

در این زمینه نیز چن و همکاران [۱۰۸] الگوریتم ابتکاری درجه تخفیف را برای حل مسئله حداکثر تأثیرگذاری پیشنهاد دادند. این الگوریتم از مدل آبشار مستقل مشتق شده است. تانگ و همکاران [۱۰۹] الگوریتمی با عنوان حداکثر تأثیر از طریق Martingales (IMM) ارائه دادند. در مدل IMM، برخی از تکنیک‌های برآورد پیشرفته مانند ابزار آماری کلاسیک و Martingales استفاده شده است. یک پارادایم دو فازی از TIM در مدل پذیرفته شده است، اما شامل یک

[1] Partitioning and Seeding

مرحله برآورد پارامتر به طور قابل توجهی متفاوت است، که قادر به استخراج یک حد پایین‌تر از OPT است، که به صورت مجانبی محکم است. تانگ و همکاران [۱۱۰] همچنین الگوریتمی به نام حداکثر سازی تأثیر دو فاز (TIM) ارائه دادند که هدف آن پل زدن تئوری و عمل در حداکثرسازی تأثیر بود. ژو و همکاران [۱۱۱] یک الگوریتم ابتکاری دو فازی (TPH) را برای حل مسئله حداکثر تأثیر مبتنی بر مکان پیشنهاد کردند. در الگوریتم، هر گره احتمال آفلاین خاص خود را برای یک محصول خاص دارد، بنابراین توجه به حداکثر سازی تأثیر مبتنی بر مکان فقط نمی تواند بر روی توپولوژی شبکه بلکه خاصیت آفلاین هر گره باشد. هو و همکاران [۱۱۲] با روشنگری الگوریتم اکتشافی تصادفی، یک روش همسایه حداکثر درجه تصادفی (RMDN) را پیشنهاد دادند.

۳.۴.۲ روش‌های دیگر

در این زمینه نیز پنگ و همکاران [۶۶] مکانیزمی برای شناسایی kـ تأثیرگذارترین گره‌ها بر اساس رأی گیری همتا طراحی کردند. هر گره با توجه به درجه صمیمیت، به تأثیرگذارترین گره در میان گره‌های دوست خود، رأی می‌دهد، سپس یک الگوریتم مرتب سازی هرمی برای مرتب سازی نتایج رأی گیری برای یافتن برترین گره‌های تأثیرگذار استفاده می‌شود. چنگ و همکاران [۱۱۳] نیز چارچوبی به نام IMRank را برای حل مسئله حداکثر سازی نفوذ تحت مدل آبشار مستقل پیشنهاد کردند. این الگوریتم با این بینش کلیدی ایجاد شد که الگوریتم‌های حریصانه اساساً یک رتبه‌بندی خود سازگار پیدا می‌کنند، جایی که درجه گره‌ها با گسترش نفوذ حاشیه‌ای مبتنی بر رتبه‌بندی آن‌ها سازگار است. همچنین چن و همکاران [۱۰۸] برای کاهش پیچیدگی محاسبات الگوریتم MixGreedyIC را پیشنهاد دادند که این الگوریتم با حذف گره‌های غیرقابل دسترس در نمودار G عملکرد را بهبود بخشیده و یک نمودار جدید کوچکتر G^* تشکیل می‌دهند. دین و همکاران [۱۱۴] هم الگوریتمی به نام محاسبه طیف نفوذ (InfSpec) برای حل مسئله تأثیر حداکثر برای همه تعداد گره‌ها برای شروع ممکن از $k = ۱$ تا n پیشنهاد دادند. همان‌گونه که بورگس و همکاران [۱۱۵] با ایجاد یک الگوریتم تقریب با فاکتور ثابت یک الگوریتم سریع برای مسئله حداکثر تأثیرگذاری فراهم کردند. لیو و همکاران [۱۱۶] یک رویکرد خطی محدود را پیشنهاد کردند که یک محدوده فوقانی محکم از تأثیر هر مجموعه گره را نشان می‌دهد.

۵.۲ جمع‌بندی

در این فصل برخی از کارهایی که در گذشته در این زمینه انجام شده‌اند را با توجه به معیارها و مدل‌های مطرح شده‌شان دسته‌بندی و بررسی کردیم. برخی از آن‌ها را به طور مفصل‌تر توضیح داده و نقات قوت و یا ایراداتشان را بیان کردیم.

فصل ۳

روش پیشنهادی

۱.۳ مقدمه

در فصل ۱ اشاره شد که هدف از این تحقیق ساخت سیستمی است که بتواند افراد تأثیرگذار در شبکه‌ی اجتماعی توییتر را شناسایی کند. همچنین با بررسی کارهای انجام شده در فصل ۲ به این نکته دست‌یافتیم که با توجه به انتزاعی بودن مفهوم تأثیر، در ابتدا باید تعریفی از تأثیر اجتماعی را ارائه داده و سپس به دنبال افراد با خصوصیات مورد نظر تحت عنوان «افراد تأثیرگذار» بگردیم. در روش پیشنهادی، بعد از تعریف مفهوم تأثیر و نفوذ اجتماعی، در مرحله‌ی اول به بررسی تأثیر هر توییت پرداخته و بعد از بدست آوردن مقدار این تأثیر، به بررسی تأثیر اشخاص در شبکه خواهیم پرداخت.

در این فصل، ابتدا تعریف مورد نظر خود از تأثیر و نفوذ افراد در شبکه‌های اجتماعی مجازی را بیان می‌کنیم، سپس برای مدل‌سازی، فاکتورهایی را که طبق روش پیشنهادی ارائه شده، بر اندازه‌ی «تأثیر»، موثراند را نیز مشخص می‌کنیم. همچنین روابطی را نشان خواهیم داد که می‌توانند برای اندازه‌گیری تأثیر اجتماعی مورد استفاده قرار گیرند. آن‌گاه با تکیه بر موارد ارائه شده به بررسی خصوصیات پایگاه داده مورد نیاز خواهیم پرداخت. در انتها نیز حسب ضرورت به بحث و نتیجه گیری موارد بدست آمده می‌پردازیم.

۲.۳ تعریف تأثیرگذاری

مفهوم نفوذ و تأثیرگذاری در شبکه‌های اجتماعی با توجه به مفاهیم ارائه شده در فصل‌های قبل، تعریف مشخص، ثابت و دقیقی ندارد و هریک از پژوهش‌های انجام شده در این زمینه، در ابتدا تعریفی را برای تأثیرگذاری ارائه داده و طبق آن تعریف راه حلی را برای مسئله‌ی مورد نظر پیشنهاد می‌کنند. بنابر این برای ارائه‌ی یک روش پیشنهادی برای تحلیل تأثیر در شبکه‌های اجتماعی، ابتدا باید تعریفی برای مفهوم نفوذ و تأثیرگذاری اجتماعی در شبکه‌های اجتماعی مجازی ارائه دهیم.

بر این اساس تأثیر گذاری را می‌توانیم به این شکل تعریف کنیم که «اگر یک فرد بتواند باعث شود فرد دیگری، کاری را انجام و یا تصمیمش در مورد موضوعی را تغییر دهد، بر روی آن فرد تأثیر گذاشته است». ما به دنبال بررسی میزان پخش شدن توییت‌ها در سطح شبکه و همچنین میزان موفقیت توییت در تغییر دادن رفتار افراد هستیم. این تأثیرگذاری می‌تواند مجاب کردن فرد برای پسندیدن (like کردن) توییت، نظر دادن (comment گذاشتن) و یا حتی تغییر عقیده‌ی فرد نسبت به یک محصول و یا یک جریان فکری ـ سیاسی و غیره باشد. بنابراین تأثیر یک توییت را به این شکل درنظر می‌گیریم که هرچه یک توییت بتواند بیشتر در شبکه پخش شود، مدت طولانی‌تری دیده شود و بازخورد دریافت کند، تأثیر بیشتری دارد.

با توجه به مطالعاتی که قبلا روی این موضوع انجام شده است و در فصل قبل نیز به آن اشاره شد، روش‌های مختلفی برای انجام این کار مورد آزمایش قرار گرفته است. مطالعاتی که در ابتدای پیدایش این موضوع انجام شدند بیشتر پارامترهایی مشابه این کار داشته‌اند، از جمله تعداد افرادی که کاربر را دنبال می‌کنند، تعداد کسانی که مطلبی که کاربر منتشر کرده را پسندیده‌اند، نظر خود را در مورد آن مطلب بیان کردند، آن مطلب را به اشتراک گذاشتند و غیره. تفاوت اصلی این کار با کارهای مشابه، درنظر گرفتن توییت‌هایی است که احتمالا برخی از مخاطبان بدون ثبت هیچ‌گونه بازخوردی، درموردشان به شکل موافق یا مخالف توییتی مستقل منتشر کرده‌اند. طبق تعریف تأثیر یک توییت، تأثیرگذاری هر فرد را نیز می‌توان میانگین تأثیر توییت‌های فرد در طول زمان دانست.

۳.۳ پارامترهای تأثیرگذار

با توجه به این فرضیات، پارامترهای مورد نیاز برای اندازه‌گیری هر توییت شامل موارد زیر خواهند بود:

- F_u : تعداد دنبال‌کنندگان کاربر منتشرکننده‌ی توییت.
- f^i_{like} : تعداد دنبال‌کنندگان کاربر i که توییت را پسندید.

- f^i_{reply} : تعداد دنبال‌کنندگان کاربر i که نظر خود را درمورد توییت بیان کردند.
- $f^i_{retweet}$: تعداد دنبال‌کنندگان کاربر i که توییت را ریتوییت کردند.
- f^i_f : تعداد دنبال‌کنندگان کاربر i که بعد از انتشار این توییت و بدون ثبت هیچ بازخوردی، اقدام به انتشار توییتی مرتبط کردند.
- F_{like} : مجموع تعداد دنبال‌کنندگان همه‌ی کسانی که توییت را پسندیدند.
- F_{reply} : مجموع تعداد دنبال‌کنندگان همه‌ی کسانی که نظر خود را درمورد توییت بیان کردند.
- $F_{retweet}$: مجموع تعداد دنبال‌کنندگان همه‌ی کسانی که توییت را ریتوییت کردند.
- F_f : مجموع تعداد دنبال‌کنندگان همه‌ی کسانی که بعد از انتشار توییت و بدون ثبت بازخورد، یک توییت مرتبط را منتشر کردند.

به این ترتیب خواهیم‌داشت:

(۱.۳)
$$F_{like} = \sum_i f^i_{like} \qquad F_{reply} = \sum_i f^i_{reply} \qquad F_{retweet} = \sum_i f^i_{retweet} \qquad F_f = \sum_i f^i_f$$

برای اندازه‌گیری میزان تأثیر هر توییت در شبکه، می‌توانیم رابطه‌ای به شکل زیر تعریف کنیم:

$$I_{tweet} = C_0 F_u + C_1 \sum_{t_0}^{t_{like}} F_{like} + C_2 \sum_{t_0}^{t_{reply}} F^i_{reply} + C_3 \sum_{t_0}^{t_{retweet}} F^i_{retweet} + C_4 \sum_{t_0}^{t_f} F^i_f \qquad (۲.۳)$$

برای محاسبه‌ی مقدار تأثیر هر کاربر نیز از میانگین تأثیر توییت‌ها استفاده می‌کنیم:

$$I_{user} = \frac{1}{N} \sum_{t=1}^{N} I_{tweet}(t) \qquad (۳.۳)$$

در رابطه‌ی ۲.۳، t_{like} ، t_{reply} ، $t_{retweet}$ و t_f زمان آخرین بازخورد دریافت شده‌ی مورد نظر و t_0 نشان‌دهنده‌ی زمان انتشار توییت است. ضرایب C_0 تا C_4 هم نشانگر میزان اهمیت هرکدام از این مؤلفه‌ها در رابطه هستند.

به طور کلی این روش از سه مرحله‌ی اساسی تشکیل شده است که در ادامه آن را شرح خواهیم‌داد.

۱. پیش‌پردازش:

- حذف stop wordها از متن توییت‌ها
- استفاده از ریشه‌ی کلمات بجای خود کلمات

۲. محاسبه میزان تأثیر هر توییت:

- یافتن توییت‌های مشابه

۳. محاسبه میزان تأثیر هر کاربر

۴.۳ پیش‌پردازش

برای یافتن توییت‌های مشابه، که عینا کپی شده‌اند و یا به میزان قابل توجهی کلمات، عبارات، کاراکترها و یا موضوعشان مشابه یکدیگر است، نیازمند بررسی متن توییت‌ها هستیم. بنابراین ابتدا باید متن توییت‌ها در یک مرحله مورد پیش‌پردازش قرار بگیرند تا عواملی که باعث کاهش دقت عملکرد این روش می‌شوند تا حد ممکن از بین بروند.

این مرحله شامل حذف stop wordها، علائم و عبارات خاصی است که احتمالا در اکثر توییت‌ها وجود داشته و اطلاعات مفیدی در مورد موضوع توییت به ما نمی‌دهند. بنابراین حذف این موارد باعث کاهش میزان محاسبات و افزایش دقت خواهد شد. پس از این کار، کلمات را به ریشه‌شان تبدیل کرد تا تفاوت بین حالت‌های مختلف یک کلمه از بین رفته و شباهت‌ها به درستی نمایان شوند. این کار با کاهش ابعاد ویژگی باعث افزایش دقت عملکرد و سرعت محاسبات می‌شود.

۵.۳ محاسبه میزان تأثیر هر توییت

میزان دیده‌شدن یک توییت به شکل قابل توجهی نشان‌دهنده‌ی مقدار تأثیر آن توییت است. وقتی شخصی توییتی را منتشر می‌کند، در مرحله‌ی اول، تمام دنبال‌کنندگان آن فرد، آن توییت را خواهند دید. در مرحله‌ی بعد، تمامی دنبال‌کنندگان همه‌ی کسانی که واکنشی به آن توییت نشان داده‌اند (توییت را پسندیدند، نظر خود را بیان کرده و یا آن توییت را ریتوییت کرده‌اند) آن توییت را خواهند دید. اما نکته‌ی مهمی که اغلب پنهان می‌ماند، این است که ممکن است شخص B توییت شخص A را دیده و با تأثیر گرفتن از آن و بدون ثبت هیچگونه بازخورد، توییتی مستقل را بدون نام بردن از شخص A منتشر کند. این توییت می‌تواند عینا همان توییت و یا صرفا از لحاظ موضوعی با آن مشابه باشد.

از این پس دنبال‌کنندگان کاربر A و دنبال‌کنندگان تمام کاربرانی که بازخوردی را برای توییت کاربر A ثبت کردند را «مخاطبان کاربر A» می‌نامیم.

برای یافتن موارد اینچنینی، باید به دنبال توییت‌های مشابهی باشیم که هم از لحاظ زمانی بعد از توییت کاربر A با یک بازه زمانی مشخص (که در این تحقیق برابر با دو ماه در نظر گرفته شده است) منتشر شده باشند و هم توسط مخاطبان کاربر A، منتشر شده باشند. بنابراین برای بدست آوردن میزان تأثیرگذاری هر توییت از کاربر A باید توییت‌های مشابه با آن را از بین مجموعه توییت‌های منتشر شده توسط مخاطبان آن کاربر را بررسی کنیم. برای پیدا کردن توییت‌های مشابه، مقایسه‌ی دو به دوی آن‌ها هزینه‌ی زیادی خواهد داشت. در صورتی که هزینه‌ی I/O، Search Query و بقیه محاسبات را در نظر نگرفته و تمام مراحل دیگر این مقایسه و محاسبات را برابر با O(1) در نظر بگیریم، آنگاه این مقایسه‌ها چیزی درحدود $O(n^۲)$ هزینه خواهد داشت.

به این منظور و برای تسریع کار و بهبود عملکرد به دلیل حجم بالای داده‌ها و همچنین محاسبات، از الگوریتم LSH [1] برای یافتن توییت‌های مشابه استفاده می‌کنیم.

الگوریتم LSH تکنیکی است که داده‌های ورودی مشابه را با احتمال بالایی به خروجی‌های مشابهی نگاشت می‌کند. در حالی که تعداد خروجی‌ها بسیار کمتر از تعداد ورودی‌ها است. این روش برای خوشه بندی و همچنین جست‌وجوی نزدیک‌ترین همسایه به کار می‌رود.

LSH

برای پیدا کردن جفت آیتم‌های مشابه در یک مجموعه‌ی بزرگ، الگوریتم LSH به ما کمک می‌کند تا بدون نیاز به بررسی همه‌ی جفت آیتم‌های ممکن با هزینه‌ی زیاد، به نتیجه‌ی مطلوب برسیم. به عنوان مثال می‌توان به یافتن جفت صفحات وب و یا سند به شکلی که متن مشترک زیادی داشته باشند (برای یافتن سرقت ادبی، Mirror sites، مقالات خبری مشابه و غیره) بعنوان یک کاربرد از این الگوریتم اشاره کرد.

این الگوریتم از سه تکنیک Shingling، Minhashing و درهم سازی بر اساس مکان (Locality sensitive hashing) استفاده می‌کند.

- **Shingling**

  تبدیل اسناد، ایمیل و غیره به مجموعه‌ها

- **Minhashing**

  تبدیل مجموعه‌های بزرگ به امضاهای کوتاه (لیستی از اعداد صحیح) به شکلی که شباهت اسناد حفظ شوند

- **Locality-Sensitive Hashing**

  انتخاب امضاهای مشابه برای بررسی اسناد مربوطه

[1] Locality Sensitive Hashing

Shingle_id	Doc(1)	Doc(2)	Doc(3)	Doc(4)
1	0	1	1	0
2	1	0	1	1
3	1	0	0	0
4	0	1	0	1
5	0	0	0	1

شکل ۱.۳: Minhashing و ایجاد ماتریس امضای اسناد

در مرحله‌ی اول، متن به شکل مجموعه‌ای از n-shingleها (n-gram) می‌شود. این عمل به این شکل انجام می‌شود که اگر بر فرض $n = ۲$ و متن مورد نظر «سلام دنیا» باشد، آن‌گاه مجموعه shingleهای مورد نظر به این شکل خواهند بود: «سل»، «لا»، «ام»، «م »، « د»، «دن»، «نی»، «یا». پس از بدست آوردن همه‌ی shingleهای همه‌ی اسناد و مرتب سازی shingleها، سندها را به شکل مجموعه‌هایی هم اندازه و به شکل ۰و۱ درمی‌آوریم که هر خانه نشان‌دهنده‌ی وجود و یا عدم وجود shingle متناظر در سند است.

با توجه به این نکته که احتمالا طول هر یک از این مجموعه‌ها بسیار بیشتر از خود سندها شده است و طول این مجموعه‌ها در میزان محاسبات و زمان محاسبات بسیار تأثیرگذار است، با استفاده از Minhashing در مرحله‌ی دوم، طول مجموعه‌ها با حفظ تشابه بین اسناد، کاهش می‌یابد. طبق این الگوریتم، شباهت بین مجموعه‌ها با معیار شباهت Jaccard سنجیده می‌شود. بنابر معیار Jaccard شباهت دو مجموعه برابر با نسبت اندازه اشتراک به اندازه اجتماع آن دو مجموعه خواهد بود (رابطه‌ی ۱.۳).

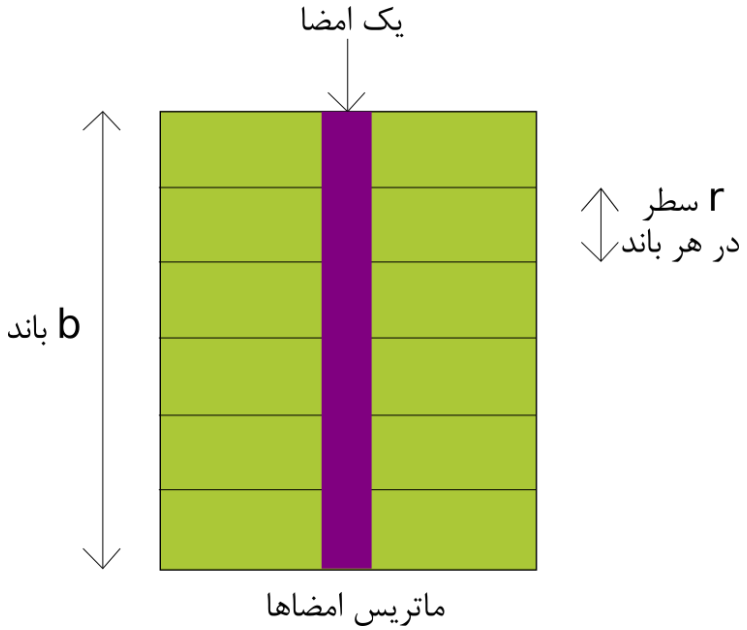
$$sim(C_۱, C_۲) = \frac{|C_۱ \cap C_۲|}{|C_۱ \cup C_۲|} \qquad (۱.۳)$$

در مرحله‌ی دوم برای تبدیل مجموعه‌ها به امضاهای متناظر با حفظ شباهت از یک تابع درهم ساز استفاده می‌شود. بعنوان مثال اگر تابع درهم ساز را به شکلی درنظر بگیریم که خروجی تابع برای هر سند، اندیس اولین ۱ موجود در مجموعه‌ی متناظر با آن سند باشد، آنگاه با تکرار این عمل با تعداد دفعات مناسب روی جایگشت‌های مختلف مجموعه‌ها، یک امضا برای هر سند بدست می‌آید که طول بسیار کمتری از آن‌ها دارد و همچنین شباهت بین امضاها به شکل قابل قبولی نزدیک به شباهت Jaccard اسناد می‌باشد.(شکل ۱.۳)

نتیجه‌ی مرحله دوم بعد از n بار تکرار با n جایگشت مختلف، امضاهایی به طول n برای هریک از مجموعه‌هاست (شکل ۲.۳). بدین ترتیب در مرحله‌ی آخر، ماتریس امضاها را به b باند تقسیم می‌کنند، به شکلی که هر باند دارای r سطر باشد.

$$r \times b = n \qquad (۲.۳)$$

سپس برای هر باند یک تابع درهم ساز در نظر گرفته و ستون‌های درون هر باند را به سطل‌هایی نگاشت می‌کنند (شکل ۳.۳). ایده این است که این کار به تعداد دفعات زیادی انجام شود تا



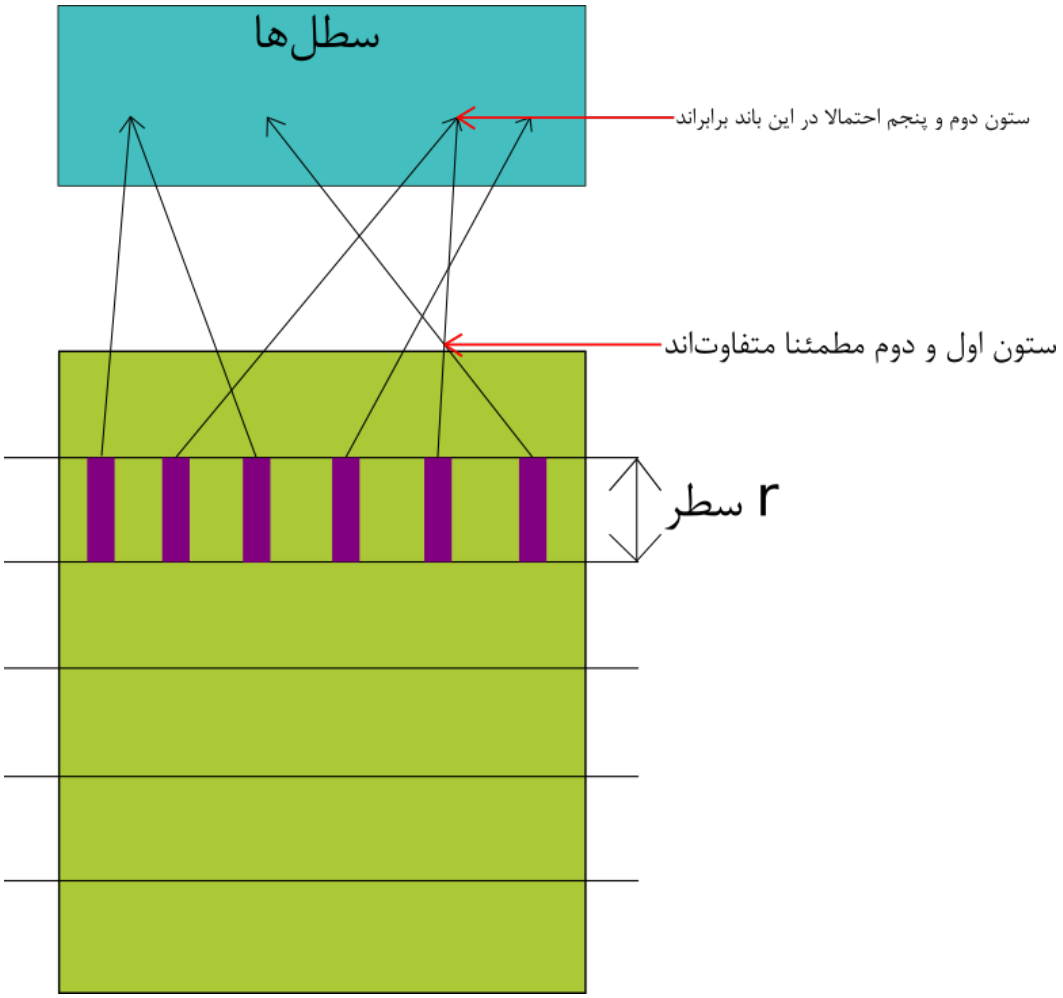
شکل ۲.۳: تقسیم‌بندی ماتریس امضاها به باندهای مختلف

مجموعه‌های شبیه به هم به سطل‌های مشترکی نگاشت شوند و همچنین امکان رخ دادن خطا حداقل شود.

با توجه به سازوکار این الگوریتم، استفاده از این روش برای کاهش هزینه‌ی زمانی و هزینه‌ی محاسبات بسیار موثر خواهد بود. بنابراین متن توییت‌ها پس از پیش‌پردازش، ابتدا در مرحله‌ی Shingling تبدیل به N-gramهایی می‌شوند و در مرحله‌ی بعد (Minhashing) امضای هر نمونه بدست می‌آید که از طول خود نمونه بسیار کوتاه‌تر است و میزان محاسبات را به خوبی کاهش می‌دهد. در مرحله‌ی آخر از روش LSH، مواردی که به شکل قابل توجهی شبیه به هم باشند با دقت خوبی در دسته‌های مشترکی قرار می‌گیرند.

۶.۳ محاسبه میزان تأثیر هر کاربر

برای محاسبه‌ی میزان تأثیر هر کاربر، از سه معیار میانگین، واریانس و تعداد استفاده می‌کنیم. معیار اصلی برای اندازه‌گیری میزان تأثیر هر کاربر در شبکه اجتماعی، میانگین مقدار تأثیر توییت‌های آن کاربر است. اما برای رتبه بندی و مقایسه‌ی تأثیر افراد با هم در صورتی که میانگین تأثیر توییت‌هایشان برابر بود، در مرحله‌ی اول از واریانس و در صورتی که واریانس‌شان نیز برابر باشد، تعداد توییت‌هایشان را در نظر می‌گیریم. بنابراین اگر بخواهیم کاربران را بر اساس میزان تأثیرشان در شبکه رتبه‌بندی کنیم، با توجه به میانگین تأثیر توییت‌هایشان آن‌ها را مرتب می‌کنیم. در صورتی که دو کاربر میانگین تأثیر توییت‌هایشان برابر باشد، تأثیر کاربری که واریانس تأثیرات توییت‌هایش کمتر باشد، بیشتر درنظر گرفته می‌شود و در صورتی که واریانس تأثیر توییت‌های این کاربران هم برابر باشد، کاربری که تعداد توییت‌هایش بیشتر



شکل ۳.۳: نگاشت امضاها به سطل‌ها با استفاده از توابع درهم‌ساز در هر باند

باشد، مؤثرتر در نظر گرفته می‌شود.

۱.۶.۳ میانگین

فرض می‌کنیم دو کاربر «آ» و «ب» را داریم به شکلی که کاربر «آ» ۵۰ توییت با میزان تأثیر ۱۰۰ و همچنین کاربر «ب» ۱۰۰۰۰ توییت با میزان تأثیر ۱ منتشر کرده باشند. اگر بخواهیم مجموع میزان تأثیر توییت‌های کاربر را بعنوان میزان تأثیر کاربر در شبکه در نظر بگیریم، آن‌گاه میزان تأثیر کاربر «آ» برابر با ۵۰۰۰ و مقدار تأثیر کاربر «ب» برابر با ۱۰۰۰۰ خواهد بود. در این صورت تعداد توییت‌ها به شکل قابل توجهی در این محاسبه‌ی این مقدار تأثیرگذار خواهد بود. بنابراین مجموع مقدار تأثیر توییت‌های کاربر نمی‌تواند معیار مناسبی برای محاسبه‌ی میزان نفوذ کاربر در شبکه‌ی اجتماعی باشد. برای حل این مشکل می‌توان از میانگین این مقادیر استفاده کرد. در صورت استفاده از میانگین برای مثال بالا مقدار تأثیر گذاری کاربر «آ» برابر با ۱۰۰ و برای کاربر «ب» برابر با ۱ خواهد بود.

۲.۶.۳ واریانس

در صورتی که مقدار میانگین تأثیرات توییت‌های دو فرد برابر باشند، برای رتبه‌بندی کاربران بر اساس میزان تأثیر، از واریانس استفاده می‌کنیم. اگر فرض کنیم میزان تأثیر توییت‌های کاربر «آ» برابر با ۲۰و۲۰و۲۰و بوده و این مقادیر برای کاربر «ب» برابر با ۱۵و ۲۰و ۲۵ باشد، در این صورت مقدار میانگین برای هر دو کاربر برابر با ۲۰ خواهد بود. در حالی که واریانس این مقادیر برای کاربر «آ» برابر با ۰ و برای کاربر «ب» برابر با ۲۵ می‌باشد. بنابراین مقدار نفوذ کاربر «آ» که تغییرات میزان تأثیر توییت‌هایش کمتر بوده‌است، بیشتر در نظر گرفته می‌شود.

۳.۶.۳ تعداد

در راستای رتبه‌بندی کاربران بر اساس میزان تأثیرشان در شبکه اجتماعی، در صورت برابر بودن مقدار میانگین تأثیرات توییت‌هایش و همچنین برابر بودن واریانس این مقادیر، کاربری که تعداد توییت‌هایش بیشتر باشد مؤثرتر درنظر گرفته می‌شود. این به این معنی‌است که آن فرد توانسته‌است با انتشار توییت‌های بیشتر، میزان تأثیر و واریانس تأثیر توییت‌هایش را حفظ کند و این به نظر می‌رسد نشانه‌ی خوبی برای تأثیر بیشتر این فرد در شبکه باشد.

۷.۳ پایگاه داده

با توجه به پارامترهای معرفی شده، پایگاه‌داده‌ی مورد نظر باید شامل موارد زیر باشد:

- **نام کاربری**

  برای ارتباط بین گراف روابط کاربران و توییت‌ها

- **تعداد دنبال‌کنندگان هر کاربر**

- **گراف رابطه‌ی کاربران (follower-following)**

- **متن توییت‌های کاربران**

  برای تشخیص شباهت بین موضوعی ناشی از تأثیر (مربوط به f_f)

- **کاربرانی که بازخورد خود را برای توییت‌ها ثبت کرده‌اند**

  کاربرانی که توییت را پسندیدند، نظر خود را بیان کردند و یا توییت را به اشتراک گذاشتند

- **زمان انتشار توییت‌ها**

- **زمان ثبت بازخوردها**

- **نوع توییت**

  - توییت[1]
  - ریتوییت[2]
  - کوت[3]

پایگاه داده‌های موجود از شبکه‌ی اجتماعی توییتر معمولا یا فقط شامل مشخصات توییت (و عمدتا فقط زمان و متن توییت) برای مقاصدی همچون تحلیل احساسات بوده و یا تنها شامل گراف کاربران و یا گراف تعاملات بین کاربران هستند. درحالی که در پایگاه داده‌ی مورد نیاز در این تحقیق اطلاعات کامل‌تری مورد نیاز است.

بعنوان مثال پایگاه داده‌ی SNAP [۱۱۷] دارای متن توییت، کاربر منتشرکننده و زمان انتشار حدود ۴۶۷ میلیون توییت است و بقیه‌ی پارامترهای مورد نیاز را شامل نمی‌شود. باقی پایگاه داده‌ها نیز هرکدام بخشی از پارامترها را داشته و باقی آن‌ها را شامل نمی‌شوند. بسیاری از این پایگاه داده‌ها ازجمله SNAP به دلیل تغییر سیاست‌های حفظ حریم شخصی توییتر و به خواست آن، دیتاست‌های خود را حذف کرده‌اند.

با توجه به این مشکلات گزینه‌های زیادی برای انتخاب موجود نبوده و ناچار باید یکی از دو راه «استفاده از API های توییتر برای جمع‌آوری داده‌های آنلاین و ساخت دیتاست» و یا «ادغام پایگاه داده‌های موجود» را انتخاب کرد. به دلیل حجم بالای کاربران و توییت‌هایی که در هر

[1] Tweet
[2] ReTweet
[3] Quote

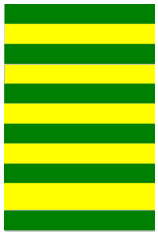
لحظه منتشر می‌شوند، عملا امکان دسترسی به همه‌ی داده‌ها وجود نداشته و علاوه بر این، توییتر هم به استفاده کنندگان از APIهای خود، اجازه‌ی ارسال تعداد محدودی درخواست را می‌دهد. بنابراین برای فقط برای جمع‌آوری تعداد قابل قبولی از توییت‌ها و همچنین دریافت گراف دنبال‌کنندگان کاربران منتشرکننده‌ی آن توییت‌ها، باید زمان بسیار زیادی را صرف کرد. بنابراین تنها راه باقی مانده ادغام پایگاه‌داده‌های موجود و استفاده از اشتراک آن‌ها است.

این محدودیت‌های ارسال درخواست به توییتر از حدود سال ۲۰۱۲ تا به امروز به دلیل حجم بالای کاربران، توییت‌ها و همچنین درخواست‌ها اعمال می‌شود. بنابراین پایگاه داده‌های جمع‌آوری شده بعد از اعمال این محدودیت‌ها همگی فقط بخشی از داده‌های واقعی هستند در حالی که پایگاه داده‌های جمع‌آوری شده قبل از اعمال محدودیت‌ها تمامی داده‌های آن زمان را شامل می‌شدند.

به این منظور از پایگاه داده‌های ArchiveTeam [۵۴، ۵۵] که حاوی مشخصات کامل هر توییت منتشر شده در لحظه‌ی انتشار در دو ماه اول سال ۲۰۱۲ بودند و همچنین از پایگاه داده‌ی twitter social graph[۵۶] که گراف کامل کاربران توییتر در سال ۲۰۱۰ بوده استفاده شده است.

دیتاست توییت‌ها، مجموعه‌ای از فایل‌های JSON و دارای مشخصات کامل ۱۶۴,۰۵۰,۰۴۵ توییت در لحظه‌ی انتشار است. بخشی از این مشخصات عبارتند از:

- geo
- retweet_count
- favorited
- text
- in_reply_to_status_id
- in_reply_to_screen_name
- in_reply_to_user_id
- retweeted
- in_reply_to_status_id
- source
- id
- entities

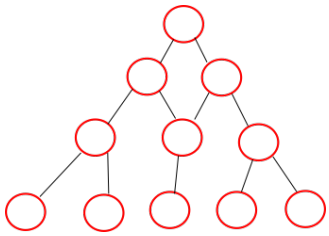


شکل ۴.۳: نمای فرضی از پایگاه داده توییت‌ها در توییتر

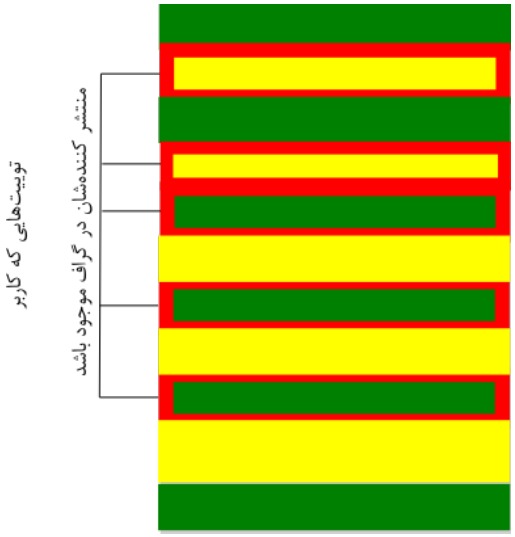
- hashtags
- urls
- user_mentions
- created_at
- user
  - followers_count
  - screen_name
  - id_str
  - created_at

گراف کاربران نیز یک فایل متنی است شامل ۱۴۶۸۳۶۵۱۸۲ سطر که هر سطر شامل دو عدد است که با فاصله از هم جدا شده‌اند. این اعداد شناسه‌ی کاربری یکتای فرد دنبال‌کننده و فرد دنبال‌شونده است. این پایگاه داده شامل ۴۱۶۵۲۲۳۰ کاربر و روابط دنبال‌کنندگی بین‌شان است. بعبارتی گراف کاربران شامل ۴۱۶۵۲۲۳۰ گره و ۱۴۶۸۳۶۵۱۸۲ یال است. نمای کلی دیتاست توییت‌ها و دیتاست گراف دنبال‌کنندگان در توییتر را به شکل۴.۳ و شکل۵.۳ در نظر می‌گیریم. برای ادغام این اطلاعات و ساخت دیتاست مورد نظر ابتدا تمام توییت‌هایی را که کاربر منتشرکننده‌ی آن در گراف کاربران وجود داشته باشد را به مجموعه‌ی توییت‌های خود اضافه می‌کنیم (۶.۳). در مرحله‌ی بعد سطرهایی از دیتاست کاربران که هم کاربر دنبال‌کننده و هم کاربر دنبال‌شونده آن حداقل یک توییت را در مجموعه توییت‌های انتخاب‌شده داشته باشند را به مجموعه کاربرهای خود اضافه کردیم.

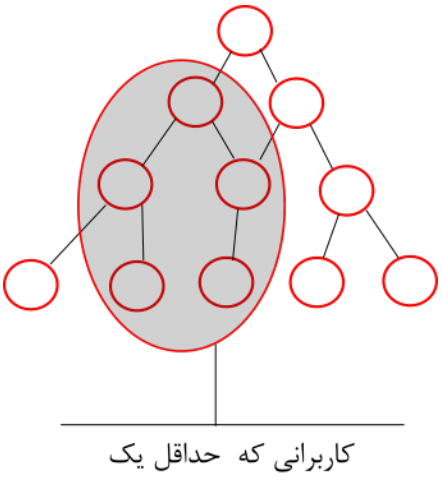
بدین ترتیب مجموعه‌ای شامل ۲۰۷۶۸۲۸ کاربر و روابط دنبال‌کنندگی بین آن‌ها و همچنین تمام توییت‌های آن‌ها به تعداد ۱۸۱۸۰۹۰۸ در دو ماه ابتدایی سال ۲۰۱۲ خواهیم بود. با وجود این داده‌ها همچنان اطلاعات پسندیدن‌ها و زمان آن‌ها را در اختیار نداریم که به ناچار ضریب



شکل ۵.۳: نمای فرضی از گراف کاربران توییتر



شکل ۶.۳: انتخاب توییت‌هایی که کاربر منتشرکننده‌ی آن را داشته باشیم



شکل ۷.۳: انتخاب کاربرانی توییتر که حداقل یک توییت از آن در مجموعه توییت‌های انتخاب‌شده وجود داشته باشد

$C_1 = 0$ درنظرمی‌گیریم. در ادامه برای پیاده‌سازی روش پیشنهادی در این تحقیق بقیه ضرایب C_2 تا C_4 را نیز برابر با ۱ درنظر گرفتیم.

۸.۳ جمع‌بندی

در این فصل روش پیشنهادی را کاملا مورد بررسی قرار دادیم. این روش برای محاسبه‌ی میزان تأثیر هر توییت، به دنبال یافتن مخاطبین آشکار و پنهان هر توییت خواهد بود و سپس با استفاده از میزان تأثیر توییت‌ها، مقدار تأثیر کاربر را محاسبه می‌کند. مجموعه مخاطبین آشکار شامل تمامی دنبال‌کنندگان کاربر منتشرکننده‌ی توییت و دنبال‌کنندگان همه‌ی کاربرانی که بازخوردی را برای این توییت به ثبت رسانده‌اند می‌شود. برای به‌دست آوردن مجموعه مخاطبان پنهان یک توییت نیز شباهت آن را با توییت‌های منتشر شده توسط مخاطبین آشکارش و بعد از زمان انتشار این توییت را بررسی کرده و در صورتی که شرایط ذکر شده در روش‌پیشنهادی برآورده می‌شدند، مخاطبین‌شان را بعنوان مخاطبین پنهان توییت مورد نظر در نظر می‌گرفتیم. برای یافتن توییت‌های مشابه از روش LSH استفاده شد. در نهایت برای محاسبه‌ی تأثیر هر کاربر از میانگین تأثیر توییت‌هایش استفاده شد.

فصل ۴

ارزیابی روش پیشنهادی

۱.۴ مقدمه

در فصل قبل، روش پیشنهادی و مراحل کار به شکل مفصل شرح داده‌شد. در این فصل می‌خواهیم عملکرد روش پیشنهادی و اجزاء آن را مورد بررسی قرار داده و با تحقیقات مشابه و نتایج‌شان مقایسه کنیم.

البته با توجه به موضوع موردبررسی و این‌که هر یک از این تحقیق‌ها تعریف متفاوتی از تأثیر اجتماعی ارائه می‌دهند و چون تعریف دقیقی از تأثیر اجتماعی وجود ندارد، نمی‌توان ادعای بهتر یا بدتر بودن یک روش در برابر روشی دیگر را عنوان کرده و درصدی از دقت عملکرد این روش‌ها را برای مقایسه ارائه داد. با این حال سعی می‌کنیم این روش را از لحاظ مختلف مورد بررسی قرار دهیم.

۲.۴ تأثیر پیش‌پردازش

در مرحله‌ی پیش‌پردازش کارهایی مانند حذف Stop Wordها، حذف علائم و اصلاح لینک‌ها، تبدیل کلمات به ریشه و غیره انجام شد. با توجه به روش پیشنهادی، برای یافتن توییت‌هایی که بدون ثبت بازخورد و با تأثیر پذیرفتن از یک توییت دیگر، منتشر شده‌اند، بدنبال یافتن توییت‌های مشابه بودیم. برای بررسی متن توییت‌ها و یافتن توییت‌های مشابه ابتدا پیش‌پردازشی

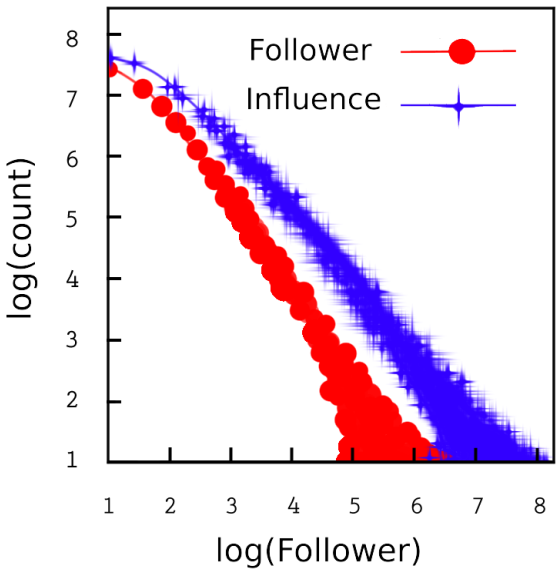
بر روی متن‌ها انجام دادیم تا توییت‌های مشابه را با دقت بیشتری پیدا کنیم. در صورتی که اگر پیش‌پردازش انجام نشود، ممکن است توییت‌هایی که واقعا مشابه نیستند، مشابه در نظر گرفته شده و توییت‌هایی که در واقع شبیه به هم هستند، مشابه در نظر گرفته نشوند. بعنوان مثال، توییت‌هایی که دارای لینک، تصویر و یا ویدئو باشند، قسمت لینک، تصویر و یا ویدئوی آن‌ها به شکل یک لینک کوتاه شده در توییتر (بعنوان مثال http://t.co/FH8Zp8Bq) نشان داده‌خواهدشد. بنابراین در صورتی که پیش پیش‌پردازش انجام نگیرد، تمامی توییت‌هایی که حاوی لینک، تصویر و یا ویدئو باشند، تا حدی شبیه به هم در نظر گرفته می‌شوند. یا بعنوان مثال در صورتی که دو توییت با متن‌های «هوا بارانی است» و «باران نمی‌بارد» داشته باشیم و برای پیش‌پردازش کلمات را به ریشه‌شان تبدیل نکنیم، «باران» و «بارانی»، دو کلمه متفاوت در نظر گرفته شده و هیچ شباهتی بینشان نخواهد بود، در صورتی که با تبدیل کلمات به ریشه، کلمه «بارانی» به «باران» تبدیل شده و نقطه اشتراک و دلیل تشابه این دو متن خواهد بود. بدین ترتیب می‌توان نتیجه گرفت که بدون انجام پیش‌پردازش، یافتن توییت‌های مشابه با خطای بسیار زیادی مواجه خواهد شد و تمام محاسبات را تحت تأثیر قرار خواهد داد.

۳.۴ بررسی عملکرد روش یافتن توییت‌های مشابه

معیارهای مختلفی از جمله معیار شباهت Jaccard و معیار شباهت کسینوسی برای محاسبه‌ی میزان شباهت بین دو متن وجود دارد. همان‌طور که در فصل ۳ عملکرد این روش به‌طور مفصل توضیح داده شد، در این تحقیق برای یافتن توییت‌های مشابه از الگوریتم LSH و معیار شباهت Jaccard استفاده شده است. معیار این انتخاب سرعت بالای الگوریتم LSH بعلت عدم مقایسه‌ی دوبه‌دوی همه‌ی موارد و کاهش هزینه‌ی زمانی بوده است.

با توجه به رتبه بندی بدست‌آمده برای کاربران از لحاظ مقدار تأثیرگذاری در شبکه نتایجی به شکل زیر بدست آمده است.

همان‌طور که در شکل ۱.۴ قابل مشاهده است، به نظر می‌رسد هر دو نمودار دنبال‌کننده و تأثیر، دارای توزیع power law [۱۱۸] هستند. با توجه به نمودار دنبال‌کنندگان واضح است که کاربران زیادی وجود دارند که تعداد دنبال‌کنندگان بسیار کمی دارند، اما تعداد کاربران با دنبال‌کنندگان زیاد، بسیار کم است. همچنین با توجه به نمودار تأثیر، می‌توان به این نکته پی برد که تعداد اندکی از کاربران شبکه‌ی اجتماعی توییتر بسیار تأثیرگذار هستند و اکثریت جامعه‌ی کاربران از افرادی تشکیل شده است که قدرت تأثیرگذاری کمی دارند. با توجه به روش پیشنهادی، در صورتی که هیچ یک از توییت‌های یک کاربر هیچ‌گونه بازخوردی دریافت نکنند، میزان تأثیر کاربر برابر با تعداد دنبال‌کنندگانش خواهد بود، بنابراین همان‌طور که در

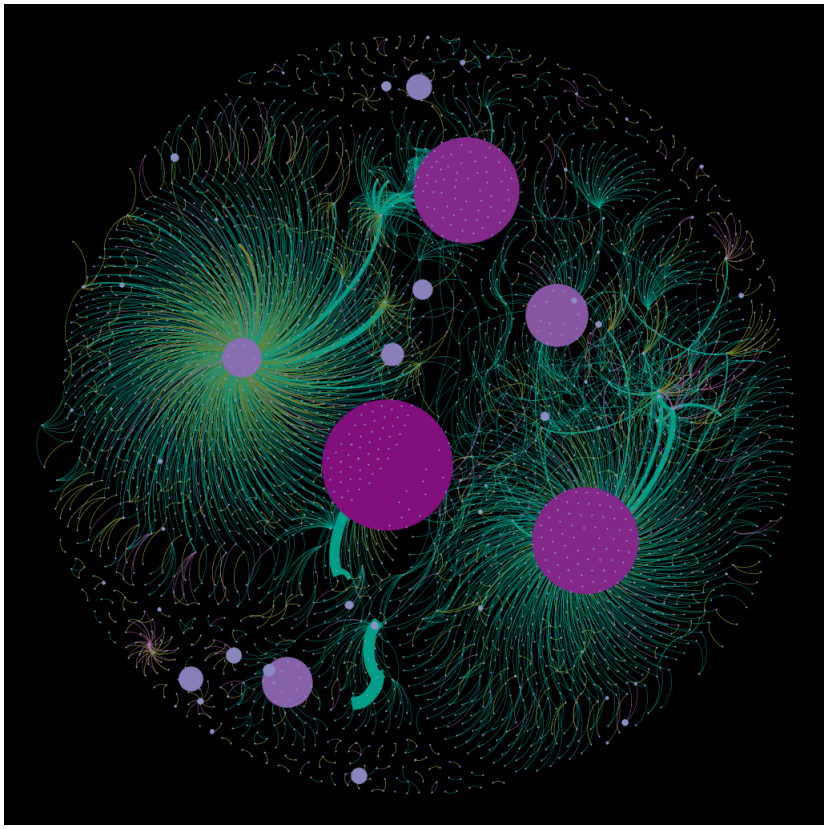


شکل ۱.۴: نمودار دنبال‌کننده ـ تأثیر

نمودار هم قابل مشاهده است، در بیشتر موارد میزان تأثیر بزرگتر از تعداد دنبال‌کنندگان و در بعضی از موارد برابر با تعداد دنبال‌کنندگان می‌باشد. همان‌طور که انتظار می‌رود، بطور معمول با افزایش تعداد دنبال‌کنندگان، میزان تأثیرگذاری هم افزایش می‌یابد، به همین دلیل هم هر دو نمودار موجود در این شکل بسیار شبیه به هم هستند. هرچند که در این بین مواردی هم یافت می‌شوند که از این قاعده پیروی نمی‌کنند. هرچه تعداد دنبال‌کنندگان افزایش می‌یابد فاصله‌ی بین دو نمودار هم بیشتر شده است و این به این معنی‌است که با افزایش دنبال‌کنندگان احتمال دریافت بازخورد بیشتر شده و طبق روش پیشنهادی، کاربران بیشتری تحت تأثیر قرار خواهند گرفت.

شکل ۲.۴ نشان‌دهنده‌ی گراف کاربران و روابط بین آن‌هاست. گره‌ها در این گراف نمایان‌گر کاربران و یال‌ها نشان‌دهنده‌ی روابط بین آن‌ها می‌باشد. این روابط می‌توانند retweet، Quote، reply و غیره باشند. یال‌های متناظر با هر یک از این روابط با رنگ جداگانه‌ای در این گراف به نمایش در آمده است.

همان‌طور که قابل مشاهده است، قطر این یال‌ها نیز ممکن است متفاوت باشند. تفاوت قطر یال‌ها نیز به معنی تفاوت در میزان رابطه بین دو کاربر و در نتیجه تفاوت تأثیر یک کاربر بر دیگری و یا به عبارتی سهم یک کاربر در افزایش تأثیرگذاری کاربر دیگر است. این حرف به این معنی‌است که اگر قطر یال بین دو گره کم باشد رابطه‌ی بین آن‌ها نیز کم است، یا کاربر اول تأثیر کمی بر روی کاربر دوم دارد و یا بعبارتی کاربر دوم سهم کمی در افزایش تأثیر کاربر اول دارد، زیرا بازخوردهای بسیار کمی را برای توییت‌های کاربر اول به ثبت رسانده و تأثیر کمی در افزایش میزان تأثیرگذاری توییت‌های کاربر اول و در نتیجه مقدار تأثیر خود کاربر اول دارد. به همین ترتیب اگر رابطه‌ی بین دو کاربر بیشتر باشد، قطر یال‌های متناظر با این روابط بیشتر



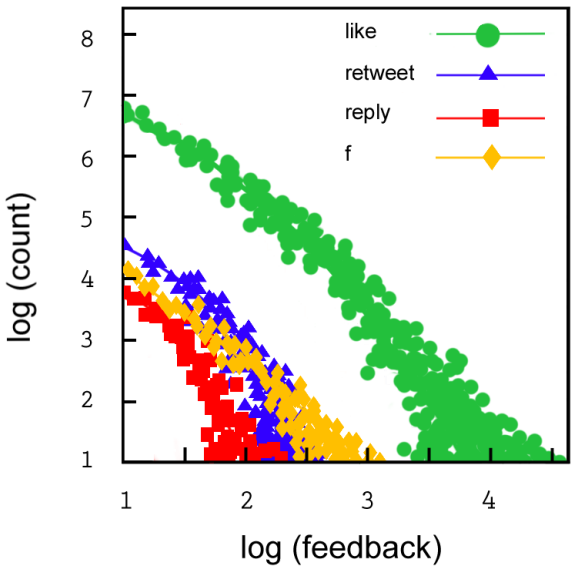
شکل ۲.۴: شبکه‌ی روابط کاربران

شده و نشان‌گر تأثیر بیشتر کاربر اول بر کاربر دوم و همچنین بیشتر بودن سهم کاربر دوم در افزایش میزان تأثیر کاربر اول خواهد بود.

رنگ گره‌ها نیز متناسب با تعداد دنبال‌کنندگان آن‌ها و اندازه هر گره نشان‌دهنده‌ی میزان تأثیرش است. همان‌طور که در نمودار شکل ۱.۴ نیز بررسی شد، در این شکل هم مشاده می‌شود که کاربران بسیار زیادی با تعداد دنبال‌کنندگان کم و همچنین تأثیر کم وجود دارند، درحالی که کاربران با تعداد دنبال‌کنندگان زیاد و تأثیر زیاد در شبکه بسیار کم تعداد هستند.

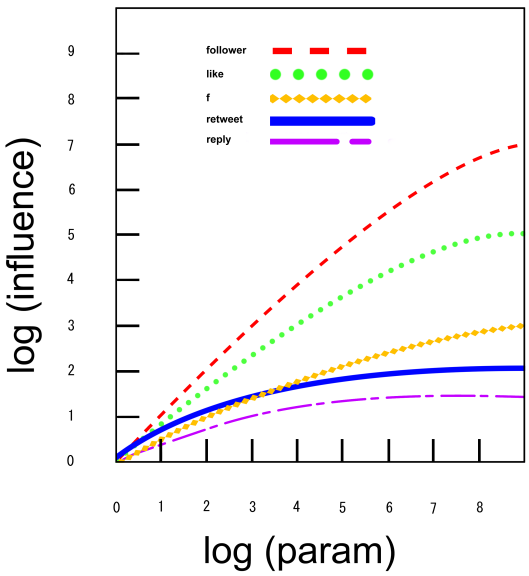
برخلاف کارهای دیگری که در این زمینه انجام شده‌اند، [۱۷] پارامترهایی که برای مدل‌سازی استفاده شده است، شامل تعداد دنبال‌کنندگان و یا تعداد بازخوردها نمی‌شود بلکه همه‌ی این پارامترها از جنس دنبال‌کنندگان هستند. بعنوان مثال همان‌طور که در رابطه‌ی ۱.۳ مشخص است، $\sum F_{like}$ مجموع تعداد دنبال‌کنندگان تمام کسانی‌است که توییت مورد نظر را پسندیده‌اند. بدین ترتیب بجای این‌که فقط به تعداد بازخوردها توجه شود، افراد و قدرت نسبی‌شان در شبکه مورد توجه قرار می‌گیرد که این موضوع می‌تواند اطلاعات واقعی‌تری را در مورد میزان تأثیر افراد به ما بدهد.

در شکل ۳.۴ نمودار مربوط به بازخوردها را مشاهده می‌کنیم. همان‌طور که پیداست، بازخوردها هم مانند تأثیر و تعداد دنبال‌کنندگان دارای توزیع power law هستند. بنابر این نمودار، تعداد کاربرانی که توییت‌شان مورد پسند واقع نمی‌شود بسیار زیاد و کاربرانی که



شکل ۳.۴: نمودار بازخورد

توییت‌هایشان توسط بسیاری از افراد مورد پسند واقع می‌شوند، بسیار کم است که برای معیارهای retweet، reply و معیار f نیز به همین ترتیب خواهد بود. نکته‌ی دیگر فاصله‌ی معنادار بین نمودار like و بازخوردهای دیگر است. بنابراین بطور کلی احتمال پسندیده شدن یک توییت بیشتر از دریافت انواع دیگر بازخورد می‌باشد.



شکل ۴.۴: نمودار مؤلفه ـ تأثیر

همان‌طور که انتظار داشتیم با افزایش مقدار هرکدام از این مؤلفه‌های ارائه شده در روش پیشنهادی، مقدار تأثیر کاربر نیز افزایش پیدا خواهد کرد. همان‌طور که در شکل ۴.۴ قابل

مشاهده است، نمودارهای مرتبط با مؤلفه‌های مختلف در برابر مقدار تأثیر به نمایش درآمده است. طبق این نتایج بدست‌آمده، مقدار مؤلفه‌ی f بعد از تعداد دنبال‌کنندگان و تعداد likeها، بیشترین اثر را در مقدار تأثیر کاربر دارد.

۴.۴ نتیجه‌گیری

در این فصل روش پیشنهادی را از جنبه‌های مختلف مورد ارزیابی قرار دادیم. تأثیر پیش‌پردازش را مشاهده و برتری روش یافتن توییت‌های مشابه را بررسی کردیم. تفاوت مؤلفه‌های به کار برده شده در روش پیشنهادی را با کارهای مشابه بیان کردیم و همچنین تأثیر مؤلفه‌های مختلف را نسبت به میزان تأثیرگذاری مشاده کرده و نشان‌دادیم که روش پیشنهادی با استفاده از مؤلفه‌های ارائه شده به خوبی قادر به محاسبه‌ی میزان تأثیرگذاری کاربران در شبکه اجتماعی توییتر می‌باشد.

فصل ۵

جمع‌بندی

۱.۵ مرور و نتیجه‌گیری

در فصل ۱ به طور مختصر در مورد شبکه‌های اجتماعی، تحلیل شبکه‌های اجتماعی و تحلیل تأثیر اجتماعی صحبت شد. هدف این تحقیق یافتن افراد تأثیرگذار با استفاده از تعریف مناسبی برای مفهوم تأثیرگذاری و همچنین انتخاب مؤلفه‌های مناسب برای محاسبه‌ی میزان تأثیر است.

در فصل ۲ نیز روش‌های مختلف برای شناسایی افراد تأثیرگذار با توجه به کاربردهای مختلف و از جنبه‌های گوناگون مورد بررسی قرار گرفت. روش پیشنهادی این تحقیق نیز به طور کامل در فصل ۳ ارائه شد. طبق روش پیشنهادی برای محاسبه‌ی میزان تأثیر هر فرد، ابتدا نیاز است تا مقدار تأثیر هر توییت محاسبه شود. به طور کلی برای محاسبه‌ی میزان تأثیر هر توییت لازم است تعداد مخاطبین توییت که بصورت بالقوه قادر به دیدن توییت هستند بدست بیاید. مخاطبین هر توییت به دو دسته‌ی مخاطبین آشکار و پنهان تقسیم‌بندی می‌شوند. مخاطبین آشکار مجموعه‌ای شامل دنبال‌کنندگان کاربر منتشرکننده‌ی توییت و دنبال‌کنندگان همه‌ی کاربرانی که بازخوردی (شامل پسندیدن، ثبت نظر، به اشتراک گذاشتن) را برای آن توییت به ثبت رسانده‌اند می‌شود. مخاطبین پنهان توییت T نیز برابر با مخاطبین آشکاه همه‌ی توییت‌هایی است که کاربر منتشرکننده‌ی آن‌ها، یکی از مخاطبین آشکار توییت T بوده و هیچ بازخوردی را نسبت به توییت T به ثبت نرسانده باشد.

برای یافتن مخاطبین پنهان نیاز است تا توییت‌هایی که از لحاظ موضوع یا محتوا مشابه توییت T هستند را بیابیم. بدین منظور بعد از پیش‌پردازش متن توییت‌ها برای یافتن موارد مشابه به جای مقایسه‌ی دو به دوی همه‌ی توییت‌ها از الگوریتم LSH استفاده شد. پس از یافتن توییت‌های مشابه و مخاطبین پنهان میزان تأثیر هر توییت محاسبه شده و سپس از میانگین تأثیر توییت‌های هر فرد مقدار تأثیر فرد بدست آمد.

در فصل ۴ روش پیشنهادی را از جنبه‌های مختلف مورد ارزیابی قرار دادیم. تأثیر پیش‌پردازش را مشاهده و برتری روش LSH یافتن توییت‌های مشابه را بررسی کردیم. تفاوت مؤلفه‌های به کار برده شده در روش پیشنهادی را با کارهای مشابه بیان کردیم و همچنین تأثیر مؤلفه‌های مختلف را نسبت به میزان تأثیرگذاری مشاده کرده و نشان‌دادیم که روش پیشنهادی با استفاده از مؤلفه‌های ارائه شده به خوبی قادر به محاسبه‌ی میزان تأثیرگذاری کاربران در شبکه اجتماعی توییتر می‌باشد.

۲.۵ کارهای آینده

با این که سعی شده است تا در روش پیشنهادی زوایای پنهان تأثیرگذاری اجتماعی هم مورد توجه قرار گیرد تا باعث افزایش دقت در یافتن افراد تأثیرگذار شود، با این حال می‌توان با مواردی که در ذیل می‌آید به اطمینان بخش تر شدن این روش افزود.

- برای این تحقیق با توجه به مدل ارائه شده، پایگاه داده‌ی مناسبی وجود نداشته، بنابر این برای ادامه‌ی تحقیق می‌توان پایگاه داده‌ی کامل‌تری را با توجه به نیازهای مدل ارائه شده جمع‌آوری کرد.
- می‌توان کاربران غیرفعال را شناسایی و حذف کرد. بعنوان مثال در صورتی که یک کاربر تعداد زیادی دنبال‌کننده‌ی غیرفعال داشته باشد، طبق روش ارائه شده، میزان تأثیر نسبتا بالایی نیز خواهد داشت که با درنظر نگرفتن کاربران غیرفعال این مشکل می‌تواند برطرف شود.
- همچنین می‌توان میزان و واریانس فعالیت کاربر در طول زمان را در محاسبه‌ی مقدار تأثیر کاربر دخیل دانست. بعنوان مثال درصورتی که یک کاربر تنها یک توییت را منتشر کرده باشد و آن توییت بصورت اتفاقی میزان بازخورد زیادی را دریافت کرده و در نتیجه تأثیر زیادی داشته باشد، مقدار تأثیر آن کاربر نیز برابر با میزان تأثیر آن توییت خواهد بود. این درصورتی است که با درنظر گرفتن میزان و واریانس فعالیت کاربر می‌توان بصورت دقیق‌تری اندازه‌ی تأثیر کاربر را محاسبه کرد.

انجام هر پژوهش علمی نیازمند مدل نظری و پایایی است که بتواند تمامیت یک تحقیق را در دامنه خود قوام داده و به شکل معقول و قابل فهمی به تصویر در آورد. هر ساختار علمی

محتاج چارچوبی است که محقق را در تدوین پروژه علمی از آغاز تا انجام هدایت می‌نماید و روند اجرای آن تسهیل واطمینان بخش می‌نماید. مهمترین این موارد بقرار زیر است: ۱ـ بیان گزاره‌های تعریف: تحقیقا بدون وجود یک تعریف مشخص و قابل فهم، انجام هر پژوهشی ناقص به نظر می‌رسد. تعاریف در دامنه یک روش علمی و قابل اندازه‌گیری معنا پیدا می‌کند. ۲ـ هر پژوهش علمی مشحون از مفاهیمی است که می‌بایست به طریق قابل فهمی توصیف گردد تا کلیت مفاهیم در یک پیوستار قابل درک، به تصویر کشیده شود. ۳ـ مضمون دیگری که حوزه‌های نظری در اختیار ما می‌گذارند محدوده‌ای است که در آن به تفحص می‌پردازیم تا از ورود متغیرهای غیر ضرور پیشگیری به عمل آورد. ۴ـ قسم دیگر کارکرد این نظریه‌ها در تدوین روش و ابزار اندازه‌گیری است که به مدد آن نتایج و یافته‌ها را نیز قابل فهم و اطمینان بخش می‌کند. بر این اساس در پژوهش حاضر و دیگر پژوهش‌های انجام گرفته، به علت نبود چنین مبنای اولیه‌ای، تعیین حدود تعاریف و دیگر اجزای تحقیق به عنوان یک دیدگاه فردی که توسط خود محقق مد نظر قرار می‌گیرد می‌تواند به دیگر محققان پیشنهاد شود. برای نیل به یک مدل علمی و قابل استناد برای دیگر پژوهش‌های در این راستا پیشنهاد می‌شود که محققان بخشی از تلاش‌های تحقیقی خود را برای یافتن مدل‌های نظری که به اتفاق نظر محققان دیگر نیز می‌انجامد معطوف نمایند.

مراجع

[1] Aggarwal, C. C. (2011) 'An Introduction to Social Network Data Analytics', in Aggarwal, C. C. (ed.) Social Network Data Analytics. Boston, MA: Springer US, pp. 1–15.

[2] Social media active users by network (October 2020) URL https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[3] Peng, S. et al. (2017) 'Social influence modeling using information theory in mobile social networks', Information Sciences, 379, pp. 146–159.

[4] Vastardis, N. and Yang, K. (2013) 'Mobile Social Networks: Architectures, Social Properties, and Key Research Challenges', IEEE Communications Surveys Tutorials, 15(3), pp. 1355–1371.

[5] Social news (December 2020).

URL https://en.wikipedia.org/wiki/Social news

[6] Kaplan, A. M. and Haenlein, M. (2011) 'The early bird catches the news: Nine things you should know about micro-blogging', Business Horizons, 54(2), pp. 105–113.

[7] Aichner, T. and Jacob, F. (2015) 'Measuring the Degree of Corporate Social Media Use', International Journal of Market Research, 57(2), pp. 257–276.

[8] Welch, M. J. et al. (2011) 'Topical semantics of twitter links', in Proceedings of the fourth ACM international conference on Web search and data mining. New York, NY, USA: Association for Computing Machinery (WSDM '11), pp. 327–336.

[9] Stumbleupon (October 2020). URL https://en.wikipedia.org/wiki/StumbleUpon

[10] Internet forum (October 2020). URL https://en.wikipedia.org/wiki/Internet forum

[11] Douban (October 2020). URL https://en.wikipedia.org/wiki/Douban

[12] Youtube (October 2020). URL https://en.wikipedia.org/wiki/YouTube

[13] Social networking service (October 2020). URL https://en.wikipedia.org/wiki/Social-net-working-service

[14] Domingos, P. and Richardson, M. (2001) 'Mining the network value of customers', in Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery (KDD '01), pp. 57–66.

[15] Cha, M. et al. (2010) 'Measuring User Influence in Twitter: The Million Follower Fallacy', Proceedings of the International AAAI Conference on Web and Social Media, 4(1), pp. 10-17.

[16] Agarwal, N. et al. (2008) 'Identifying the influential bloggers in a community', in Proceedings of the 2008 International Conference on Web Search and Data Mining. New York, NY, USA: Association for Computing Machinery (WSDM '08), pp. 207–218.

[17] Bakshy, E. et al. (2011) 'Everyone's an influencer: quantifying influence on twitter', in Proceedings of the fourth ACM international conference on Web search and data mining. New York, NY, USA: Association for Computing Machinery (WSDM '11), pp. 65–74.

[18] Huang, J. et al. (2012) 'Exploring social influence via posterior effect of word-of-mouth recommendations', in Proceedings of the fifth ACM international conference on Web search and data mining. New York, NY, USA: Association for Computing Machinery (WSDM '12), pp. 573–582.

[19] Tang, X. and Yang, C. C. (2012) 'Ranking User Influence in Healthcare Social Media', ACM Transactions on Intelligent Systems and Technology, 3(4), p. 73:1-73:21.

[20] Phan, N. et al. (2016) 'Topic-Aware Physical Activity Propagation in a Health Social Network', IEEE Intelligent Systems, 31(1), pp. 5–14.

[21] Wang, G. et al. (2014) 'Fine-Grained Feature-Based Social Influence Evaluation in Online Social Networks', IEEE Transactions on Parallel and Distributed Systems, 25(9), pp. 2286–2296.

[22] Sun, Jianshan et al. (2015) 'Leverage RAF to find domain experts on research social network services: A big data analytics methodology with MapReduce framework', International Journal of Production Economics, 165, pp. 185–193.

[23] Sang, J. and Xu, C. (2013) 'Social influence analysis and application on multimedia sharing websites', ACM Transactions on Multimedia Computing, Communications, and Applications, 9(1s), pp. 1–24.

[24] Dietz, L., Bickel, S. and Scheffer, T. (2007) 'Unsupervised prediction of citation influences', in Proceedings of the 24th international conference on Machine learning - ICML '07. the 24th international conference, Corvalis, Oregon: ACM Press, pp. 233–240.

[25] Guo, Z. et al. (2014) 'A Two-Level Topic Model Towards Knowledge Discovery from Citation Networks', IEEE Transactions on Knowledge and Data Engineering, 26(4), pp. 780–794.

[26] Hamidian, S. and Diab, M. T. (2019) 'Rumor Detection and Classification for Twitter Data', arXiv:1912.08926.

[27] Liu, K.-L., Li, W.-J. and Guo, M. (2012) 'Emoticon smoothed language models for twitter sentiment analysis', in Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. Toronto, Ontario, Canada: AAAI Press (AAAI'12), pp. 1678–1684.

[28] Peng, S., Wang, G. and Xie, D. (2017) 'Social Influence Analysis in Social Networking Big Data: Opportunities and Challenges', IEEE Network, 31(1), pp. 11–17.

[29] Yu, S. et al. (2017) 'Networking for Big Data: A Survey', IEEE Communications Surveys Tutorials, 19(1), pp. 531–549.

[30] Tang, J. et al. (2009) 'Social influence analysis in large-scale networks', in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery (KDD '09), pp. 807–816.

[31] Rabade, R., Mishra, N. and Sharma, S. (2014) 'Survey of Influential User Identification Techniques in Online Social Networks', in Thampi, S. M. et al. (eds) Recent Advances in Intelligent Informatics. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing), pp. 359–370.

[32] Batrinca, B. and Treleaven, P. C. (2015) 'Social media analytics: a survey of techniques, tools and platforms', AI and SOCIETY, 30(1), pp. 89–116.

[33] Sun, J. and Tang, J. (2011) 'A Survey of Models and Algorithms for Social Influence Analysis', in Aggarwal, C. C. (ed.) Social Network Data Analytics. Boston, MA: Springer US, pp. 177–214.

[34] Kempe, D., Kleinberg, J. and Tardos, É. (2003) 'Maximizing the spread of influence through a social network', in Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery (KDD '03), pp. 137–146.

[35] Koch, J. et al. (2016) 'An empirical comparison of Big Graph frameworks in the context of network analysis', Social Network Analysis and Mining, 6(1), p. 84.

[36] Peng, S. et al. (2017) 'Social influence modeling using information theory in mobile social networks', Information Sciences, 379, pp. 146–159.

[37] Cercel, D.-C. and Trausan-Matu, S. (2014) 'Opinion Propagation in Online Social Networks: A Survey', in Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14). New York, NY, USA: Association for Computing Machinery (WIMS '14), pp. 1–10.

[38] Peng, S. et al. (2014) 'Containing smartphone worm propagation with an influence maximization algorithm', Computer Networks, 74, pp. 103–113.

[39] Hu, J. et al. (2014) 'Analysis of influence maximization in large-scale social networks', ACM SIGMETRICS Performance Evaluation Review, 41(4), pp. 78–81.

[40] Kayes, I. et al. (2012) 'How Influential Are You: Detecting Influential Bloggers in a Blogging Community', in Aberer, K. et al. (eds) Social Informatics. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 29–42.

[41] Jabeur, L. B., Tamine, L. and Boughanem, M. (2012) 'Featured Tweet Search: Modeling Time and Social Influence for Microblog Retrieval', in 2012 IEEE/WIC/ACM

International Conferences on Web Intelligence and Intelligent Agent Technology. 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pp. 166–173.

[42] Weimann, G. (1994) The Influentials: People Who Influence People. SUNY Press.

[43] Lu, C.-T., Shuai, H.-H. and Yu, P. S. (2014) 'Identifying Your Customers in Social Networks', in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York, NY, USA: Association for Computing Machinery (CIKM '14), pp. 391–400.

[44] Li, D. et al. (2012) 'Mining topic-level opinion influence in microblog', in Proceedings of the 21st ACM international conference on Information and knowledge management. New York, NY, USA: Association for Computing Machinery (CIKM '12), pp. 1562–1566.

[45] Li, N. and Gillet, D. (2013) 'Identifying influential scholars in academic social media platforms', in Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. New York, NY, USA: Association for Computing Machinery (ASONAM '13), pp. 608–614.

[46] Ye, M., Liu, X. and Lee, W.-C. (2012) 'Exploring social influence for recommendation: a generative model approach', in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: Association for Computing Machinery (SIGIR '12), pp. 671–680.

[47] Yang, W., Wang, H. and Yao, Y. (2015) 'An immunization strategy for social network worms based on network vertex influence', China Communications, 12(7), pp. 154–166.

[48] Sathanur, A. V. and Jandhyala, V. (2014) 'An activity-based information-theoretic annotation of social graphs', in Proceedings of the 2014 ACM conference on Web science. New York, NY, USA: Association for Computing Machinery (WebSci '14), pp. 187–191.

[49] Kourtellis, N. et al. (2013) 'Identifying high betweenness centrality nodes in large social networks', Social Network Analysis and Mining, 3(4), pp. 899–914.

[50] Song, X. et al. (2007) 'Identifying opinion leaders in the blogosphere', in Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. New York, NY, USA: Association for Computing Machinery (CIKM '07), pp. 971–974.

[51] Peng, S., Yu, S. and Yang, A. (2014) 'Smartphone Malware and Its Propagation Modeling: A Survey', IEEE Communications Surveys Tutorials, 16(2), pp. 925–941.

[52] Shang, X. et al. (2012) 'Factor Analysis for Influence Maximization Problem in Social Networks', in 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, pp. 95–101.

[53] Alalwan, A. A. (2018) 'Investigating the impact of social media advertising features on customer purchase intention', International Journal of Information Management, 42, pp. 65–77.

[54] ArchiveTeam-2012-01 (October 2020). URL https://archive.org/details/archiveteam-twitter-stream-2012-01

[55] ArchiveTeam-2012-02 (October 2020). URL https://archive.org/details/archiveteam-twitter-stream-2012-02

[56] twitter social graph (October 2020). URL http://an.kaist.ac.kr/traces/WWW2010.html

[57] Okamoto, K., Chen, W. and Li, X.-Y. (2008) 'Ranking of Closeness Centrality for Large-Scale Social Networks', in Preparata, F. P., Wu, X., and Yin, J. (eds) Frontiers in Algorithmics. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 186–195.

[58] Kleinberg, J. M. (1999) 'Authoritative sources in a hyperlinked environment', Journal of the ACM, 46(5), pp. 604–632.

[59] Brin, S. and Page, L. (1998) 'The Anatomy of a Large-Scale Hypertextual Web Search Engine', Computer Networks, 30, pp. 107–117.

[60] Peng, S., Li, J. and Yang, A. (2015) 'Entropy-Based Social Influence Evaluation in Mobile Social Networks', in Wang, G. et al. (eds) Algorithms and Architectures

for Parallel Processing. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 637–647.

[61] Sun, B. and Ng, V. T. (2013) 'Identifying Influential Users by Their Postings in Social Networks', in Atzmueller, M. et al. (eds) Ubiquitous Social Media Analysis. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 128–151.

[62] Ver Steeg, G. and Galstyan, A. (2013) 'Information-theoretic measures of influence based on content dynamics', in Proceedings of the sixth ACM international conference on Web search and data mining. New York, NY, USA: Association for Computing Machinery (WSDM '13), pp. 3–12.

[63] He, S. et al. (2013) 'Identifying Peer Influence in Online Social Networks Using Transfer Entropy', in Wang, G. A. et al. (eds) Intelligence and Security Informatics. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 47–61.

[64] Chen, W. et al. (2012) 'InfluenceRank: An Efficient Social Influence Measurement for Millions of Users in Microblog', in 2012 Second International Conference on Cloud and Green Computing. 2012 Second International Conference on Cloud and Green Computing, pp. 563–570.

[65] Hui, P. and Gregory, M. (2010) 'Quantifying sentiment and influence in blogspaces', in Proceedings of the First Workshop on Social Media Analytics. New York, NY, USA: Association for Computing Machinery (SOMA '10), pp. 53–61.

[66] Peng, S., Wang, G. and Yu, S. (2013) 'Mining Mechanism of Top-k Influential Nodes Based on Voting Algorithm in Mobile Social Networks', in 2013 IEEE 10th International Conference on High Performance Computing and Communications 2013 IEEE International Conference on Embedded and Ubiquitous Computing. 2013 IEEE 10th International Conference on High Performance Computing and Communications 2013 IEEE International Conference on Embedded and Ubiquitous Computing, pp. 2194–2199.

[67] Peng, S. et al. (2019) 'An Immunization Framework for Social Networks Through Big Data Based Influence Modeling', IEEE Transactions on Dependable and Secure Computing, 16(6), pp. 984–995.

[68] Ding, Z. et al. (2013) 'An Influence Strength Measurement via Time-Aware Probabilistic Generative Model for Microblogs', in Ishikawa, Y. et al. (eds) Web Tech-

nologies and Applications. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 372–383.

[69] Iwata, T., Shah, A. and Ghahramani, Z. (2013) 'Discovering latent influence in online social activities via shared cascade poisson processes', in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery (KDD '13), pp. 266–274.

[70] Zhaoyun, D. et al. (2013) 'Mining topical influencers based on the multi-relational network in micro-blogging sites', China Communications, 10(1), pp. 93–104.

[71] Xiang, R., Neville, J. and Rogati, M. (2010) 'Modeling relationship strength in online social networks', in Proceedings of the 19th international conference on World wide web. New York, NY, USA: Association for Computing Machinery (WWW '10), pp. 981–990.

[72] Pal, A. and Counts, S. (2011) 'Identifying topical authorities in microblogs', in Proceedings of the fourth ACM international conference on Web search and data mining. New York, NY, USA: Association for Computing Machinery (WSDM '11), pp. 45–54.

[73] Kwak, H. et al. (2010) 'What is Twitter, a social network or a news media?', in Proceedings of the 19th international conference on World wide web. New York, NY, USA: Association for Computing Machinery (WWW '10), pp. 591–600.

[74] Goyal, A., Bonchi, F. and Lakshmanan, L. V. S. (2010) 'Learning influence probabilities in social networks', in Proceedings of the third ACM international conference on Web search and data mining. New York, NY, USA: Association for Computing Machinery (WSDM '10), pp. 241–250.

[75] Hajian, B. and White, T. (2011) 'Modelling Influence in a Social Network: Metrics and Evaluation', in 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing. 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 497–500.

[76] Bi, B. et al. (2014) 'Scalable topic-specific influence analysis on microblogs', in Proceedings of the 7th ACM international conference on Web search and data mining.

New York, NY, USA: Association for Computing Machinery (WSDM '14), pp. 513–522.

[77] Gerrish, S. M. and Blei, D. M. (2010) 'A language-based approach to measuring scholarly impact', in Proceedings of the 27th International Conference on International Conference on Machine Learning. Madison, WI, USA: Omnipress (ICML'10), pp. 375–382.

[78] Tang, J., Wu, S. and Sun, J. (2013) 'Confluence: conformity influence in large social networks', in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery (KDD '13), pp. 347–355.

[79] Herzig, J., Mass, Y. and Roitman, H. (2014) 'An author-reader influence model for detecting topic-based influencers in social media', in Proceedings of the 25th ACM conference on Hypertext and social media. New York, NY, USA: Association for Computing Machinery (HT '14), pp. 46–55.

[80] Zhang, J. et al. (2014) 'Role-Aware Conformity Modeling and Analysis in Social Networks', Proceedings of the AAAI Conference on Artificial Intelligence, 28(1).

[81] Cui, P. et al. (2011) 'Who should share what? item-level social influence prediction for users and posts ranking', in Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. New York, NY, USA: Association for Computing Machinery (SIGIR '11), pp. 185–194.

[82] Li, J. et al. (2016) 'Personalized Influential Topic Search via Social Network Summarization', IEEE Transactions on Knowledge and Data Engineering, 28(7), pp. 1820–1834.

[83] Zi Wang, Shinkuma, R. and Takahashi, T. (2015) 'Dynamic social influence modeling from perspective of gray-scale mixing process', in 2015 Eighth International Conference on Mobile Computing and Ubiquitous Networking (ICMU). 2015 Eighth International Conference on Mobile Computing and Ubiquitous Networking (ICMU), pp. 1–6.

[84] Liu, L. et al. (2012) 'Learning influence from heterogeneous social networks', Data Mining and Knowledge Discovery, 25(3), pp. 511–544.

[85] Aral, S. and Walker, D. (2012) 'Identifying Influential and Susceptible Members of Social Networks', Science, 337(6092), pp. 337–341.

[86] Tang, S. et al. (2011) 'Relationship classification in large scale online social networks and its impact on information propagation', in 2011 Proceedings IEEE INFOCOM. 2011 Proceedings IEEE INFOCOM, pp. 2291–2299.

[87] Deng, Q. and Dai, Y. (2012) 'How Your Friends Influence You: Quantifying Pairwise Influences on Twitter', in 2012 International Conference on Cloud and Service Computing. 2012 International Conference on Cloud and Service Computing, pp. 185–192.

[88] Wang, C. et al. (2011) 'Dynamic Social Influence Analysis through Time-Dependent Factor Graphs', in 2011 International Conference on Advances in Social Networks Analysis and Mining. 2011 International Conference on Advances in Social Networks Analysis and Mining, pp. 239–246.

[89] Saez-Trumper, D. et al. (2012) 'Finding trendsetters in information networks', in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery (KDD '12), pp. 1014–1022.

[90] Pan, W. et al. (2012) 'Modeling Dynamical Influence in Human Interaction: Using data to make better inferences about influence within social systems', IEEE Signal Processing Magazine, 29(2), pp. 77–86.

[91] Crandall, D. et al. (2008) 'Feedback effects between similarity and social influence in online communities', in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery (KDD '08), pp. 160–168.

[92] Li, D. et al. (2017) 'Positive influence maximization in signed social networks based on simulated annealing', Neurocomputing, 260, pp. 69–78.

[93] Yang, L., Giua, A. and Li, Z. (2017) 'Minimizing the Influence Propagation in Social Networks for Linear Threshold Models', IFAC-PapersOnLine, 50(1), pp. 14465–14470.

[94] Heuristic (computer science) (October 2020). URL https://en.wikipedia.org/wiki/Heuristic (computer science)

[95] Wang, Y. and Feng, X. (2009) 'A Potential-Based Node Selection Strategy for Influence Maximization in a Social Network', in Huang, R. et al. (eds) Advanced Data Mining and Applications. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 350–361.

[96] Li, H. et al. (2015) 'Conformity-aware influence maximization in online social networks', The VLDB Journal, 24(1), pp. 117–141.

[97] Wang, Y. et al. (2010) 'Community-based greedy algorithm for mining top-K influential nodes in mobile social networks', in Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery (KDD '10), pp. 1039–1048.

[98] Zhang, Y. et al. (2010) 'Estimate on Expectation for Influence Maximization in Social Networks', in Zaki, M. J. et al. (eds) Advances in Knowledge Discovery and Data Mining. Berlin, Heidelberg: Springer (Lecture Notes in Computer Science), pp. 99–106.

[99] Liu, X. et al. (2014) 'IMGPU: GPU-Accelerated Influence Maximization in Large-Scale Social Networks', IEEE Transactions on Parallel and Distributed Systems, 25(1), pp. 136–145.

[100] Ma, H. et al. (2008) 'Mining social networks using heat diffusion processes for marketing candidates selection', in Proceedings of the 17th ACM conference on Information and knowledge management. New York, NY, USA: Association for Computing Machinery (CIKM '08), pp. 233–242.

[101] Zhou, J., Zhang, Y. and Cheng, J. (2014) 'Preference-based mining of top-K influential nodes in social networks', Future Generation Computer Systems, 31, pp. 40–47.

[102] Estevez, P. A., Vera, P. and Saito, K. (2007) 'Selecting the Most Influential Nodes in Social Networks', in 2007 International Joint Conference on Neural Networks. pp. 2397–2402.

[103] Goyal, A., Lu, W. and Lakshmanan, L. V. S. (2011) 'SIMPATH: An Efficient Algorithm for Influence Maximization under the Linear Threshold Model', in 2011 IEEE 11th International Conference on Data Mining. pp. 211–220.

[104] Lee, J.-R. and Chung, C.-W. (2014) 'A fast approximation for influence maximization in large social networks', in Proceedings of the 23rd International Conference on World Wide Web. New York, NY, USA: Association for Computing Machinery (WWW '14 Companion), pp. 1157–1162.

[105] Ma, Q. and Ma, J. (2014) 'An Efficient Influence Maximization Algorithm to Discover Influential Users in Micro-blog', in Li, F. et al. (eds) Web-Age Information Management. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 113–124.

[106] Zhang, H. et al. (2016) 'Least Cost Influence Maximization Across Multiple Social Networks', IEEE/ACM Transactions on Networking, 24(2), pp. 929–939.

[107] Ok, J. et al. (2014) 'On maximizing diffusion speed in social networks: impact of random seeding and clustering', in The 2014 ACM international conference on Measurement and modeling of computer systems. New York, NY, USA: Association for Computing Machinery (SIGMETRICS '14), pp. 301–313.

[108] Chen, W., Wang, Y. and Yang, S. (2009) 'Efficient influence maximization in social networks', in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: Association for Computing Machinery (KDD '09), pp. 199–208.

[109] Tang, Y., Shi, Y. and Xiao, X. (2015) 'Influence Maximization in Near-Linear Time: A Martingale Approach', in Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data. New York, NY, USA: Association for Computing Machinery (SIGMOD '15), pp. 1539–1554.

[110] Tang, Y., Xiao, X. and Shi, Y. (2014) 'Influence maximization: near-optimal time complexity meets practical efficiency', in Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data. New York, NY, USA: Association for Computing Machinery (SIGMOD '14), pp. 75–86.

[111] Zhou, T. et al. (2015) 'Location-Based Influence Maximization in Social Networks', in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York, NY, USA: Association for Computing Machinery (CIKM '15), pp. 1211–1220.

[112] Hu, Q. et al. (2015) 'RMDN: New Approach to Maximize Influence Spread', in 2015 IEEE 39th Annual Computer Software and Applications Conference. 2015 IEEE 39th Annual Computer Software and Applications Conference, pp. 702–711.

[113] Cheng, S. et al. (2014) 'IMRank: influence maximization via finding self-consistent ranking', in Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: Association for Computing Machinery (SIGIR '14), pp. 475–484.

[114] Dinh, T. et al. (2015) 'Social Influence Spectrum with Guarantees: Computing More in Less Time', in Thai, M. T., Nguyen, N. P., and Shen, H. (eds) Computational Social Networks. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 84–103.

[115] Borgs, C. et al. (2013) 'Maximizing Social Influence in Nearly Optimal Time', in Proceedings of the 2014 Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics (Proceedings), pp. 946–957.

[116] Liu, Q. et al. (2014) 'Influence Maximization over Large-Scale Social Networks: A Bounded Linear Approach', in Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York, NY, USA: Association for Computing Machinery (CIKM '14), pp. 171–180.

[117] SNAP (October 2020). URL https://snap.stanford.edu/data/twitter7.html

[118] Power-law probability distributions (October 2020). URL https://en.wikipedia.org/wiki/Power law

Abstract

Nowadays, identifying influential users in social networks has become one of the most popular topics. The popularity of this issue is related to various reasons such as cost reduction and targeted advertising. However, there are many challenges for researchers in this field. These challenges include the large number of users in social networks and their activities, how to define social influence, selecting appropriate metrics to measure the influence and limited access to information due to privacy of users in social networks.

The purpose of this dissertation is to find influential people in Twitter by providing a suitable definition for the concept of influence and also selecting the most appropriate metrics to measure the influence.

In the proposed method for calculating the influence of each user, first the influence of each tweet is calculated. The effect of each tweet in this way is equal to the number of people who can potentially view that tweet. This requires identifying the overt and covert audience of each tweet.

We search for similar Tweets to check if the user was affected by someone else's tweets or not. To find similar tweets, the LSH algorithm is used to reduce the amount of calculations and time cost due to the high volume of data. According to the proposed method, the effect of each user is equal to the average effect of that user's tweets. But by comparison, to rank users based on their influence if the average influence of their tweets is equal, the amount of user influence is more if the variance of the effect of their tweets is less. Also, if both the mean and variance are equal for both users, the user with more tweets will be more effective.

Due to the proposed method and the required metrics, as well as the lack of a suitable database and the impossibility of data collection, appropriate and relevant data were extracted and used from the sharing of two databases ArchiveTeam and Twitter Social Graph. Finally, according to the results, the proposed method is able to calculate the users' influence more accurately and realistically, using metrics that are more efficient than the previous ones.

Keywords: Social Networks, Influential, LSH, Twitter

Faculty of Computer Engineering

MSc Thesis in Artificial Intelligence

Discovering influential users on Twitter

By: Ali Jafari Balalami

Supervisor:

Hamid Hassanpour

Advisor:

Bagher Rahimpour Cami

February 2021