بهبود مدل های مبتنی بر ویژگی‌های پنهان پایگاه‌های دانش با استفاده از قوانین انجمنی

تعریف مساله و هدف و ضرورت

در سال‌های اخیر شاهد رشد بسیار زیادی در شبکه‌های اجتماعی بوده‌ایم و مطالعات زیادی روی این شبکه‌ها انجام گرفته است. داده‌های شبکه‌های اجتماعی یکی از ابزار محبوب برای مدل کردن رابطه و رفتار افراد و جامعه یا گروهی که در آن عضو هستن بشمار می‌رود. این داده‌ها معمولا به صورت گرافی نمایش داده می‌شود که در آن گره‌ها افراد و لبه‌ها روابط بین این افراد می‌باشد. پیش‌بینی وقوع لینک‌ها، یک مساله اساسی و بنیادین در شبکه‌های اجتماعی می‌باشد. در موضوع پیش‌بینی لینک، نمایی از یک شبکه به ما داده می‌شود و ما مایل هستیم که بدانیم در آینده نزدیک، احتمالا چه تراکنش‌هایی میان اعضای فعلی شبکه روی می‌دهد و یا اینکه کدام یک از تراکنش‌های موجود را از دست می‌دهیم.

اکثر مطالعات انجام شده در این زمینه روی شبکه‌های تک‌رابطه‌ای بوده است. به این معنا که روابط بین موجودیت‌ها از یک نوع است و این روابط به صورت دوتایی‌های مرتب مورد استفاده قرار می‌گیرند. برای مثال اگر در یک شبکه اجتماعی رابطه را دوستی بین افراد در نظر بگیریم یال‌های گراف شبکه به صورت «الف، ب» خواهد بود به این معنی که شخص الف با شخص ب رابطه دوستی دارد. در سال‌های اخیر پایگاه‌های دانشی ایجاد شده‌اند که داده‌ها در آن به صورت چند رابطه‌ای ذخیره شده‌اند و اطلاعات بیشتری از یک رابطه دوتایی بلی یا خیر به ما می‌دهند. ما در این تحقیق از داده‌های چند رابطه‌ای استفاده می‌کنیم. منظور از داده‌های چند رابطه‌ای گراف جهت‌داری است متشکل از موجودیت‌ها و روابط بین آن‌ها که بصورت «مبدا h، رابطه r، مقصد t» نمایش داده می‌شود، به این معنی که یک رابطه r بین موجودیت‌های h و t وجود دارد. برای مثال سه‌تایی «تهران، واقع در، ایران» این اطلاع را به ما می‌دهد که استان تهران داخل کشور ایران قرار دارد. در این نوع پایگاه دانش هم انواع مختلف موجودیت وجود دارد و هم انواع مختلف رابطه بین موجودیت‌ها. پایگاه‌های دانشی مانندFreeBase، Google Knowledge Graph وWordNet وجود دارند که شامل تعداد زیادی نمونه چندرابطه‌ای می‌باشند و تعداد زیادی موجودیت و روابط بین آن‌ها را می‌توان در آن‌ها یافت و از آن برای یادگیری مدل استفاده کرد. شناخت پایگاه‌دانش و دریافت اطلاعات از آن نیز یکی از مسایل پیشبینی‌لینک روی داده‌های چند رابطه‌ای است. پیشبینی لینک مقیاس‌پذیر روی داده‌های چند رابطه‌ای عمدتا با استفاده از ویژگی‌های پنهان صورت می‌گیرد. مشکل این روش‌ها تعداد بسیار زیاد پارامترها در فاز یادگیری مدل است به صورتی که انجام عملیات یادگیری مدل را غیرممکن می‌کند. کاهش تعداد پارامترهای مساله نیز که در برخی روش‌ها پیشنهاد شده نیز برخی از الگوی‌های موجود در پایگاه دانش را در نظر نمی‌گیرند و باعث کاهش دقت مدل خواهند شد. در این تمرین قصد داریم که بدون افزایش تعداد پارامترهای روش‌های الگوی پنهان شده و با استفاده از قوانین انجمنی موجود در پایگاه‌دانش عمل‌کرد این روش‌ها را بهبود بخشیم.

در این تحقیق قصد داریم که با کار روی داده‌های چند رابطه‌ای به سؤالاتی که از پایگاه‌دانش می‌شود پاسخ دهیم، به این صورت که یک مدل از روی داده‌های پایگاه دانش ساخته و آموزش داده می‌شود و پس از آن سؤالات به گونه‌ای که مقصد یا مبدأ آن مجهول باشد از مدل ایجاد شده پرسیده می‌شود. برای نمونه در مثال بالا اگر قسمت کشور مجهول باشد از پایگاه دانش پرسیده می‌شود «تهران، واقع در، ؟» و انتظار می‌رود که مدل آموزش دیده پاسخ سؤال را با دقت نسبتاً خوبی بیابد.

نتایج این مساله کاربردهای زیادی خواهد داشت به عنوان مثال در ادامه تعدادی از این کاربردها را مطرح خواهیم کرد:

- پیشبینی لینک‌های احتمالی در شبکه‌های اجتماعی، برای مثال در شبکه‌های اجتماعی بین کاربران و مطالب ثبت شده، نظرات و ... بررسی شود که روابطی مانند دوستی، پسندیدن و نپسندیدن، روابط فامیلی و... وجود دارد یا خیر
- استفاده به عنوان سیستم‌های توصیه‌گر، برای مثال کاربران و کالاها یا اشیاء موجودیت‌ها هستند و خریدن، امتیاز دادن، بررسی کردن و ... رابطه‌ها هستند که می‌توان از بررسی این روابط و موجودیت‌ها اطلاعاتی نظیر کاربرx کالایy را خواهد خرید یا خیر یا اینکه کالایx به تعداد بالا فروش خواهد رفت یا خیر بدست آورد.
- کامل‌تر کردن پایگاه‌های داده، از نتایج این تحقیق می‌توان استفاده کرد و با اضافه کردن رابطه‌های انحمنی بدست آمده به پایگاه‌داده آن را کامل‌تر کرد.

روش‌ها:

داده‌هایی که قصد کار روی آن‌ها را داریم معمولاً ابعاد بالایی خواهند داشت. برای مثال پایگاه دانشFreebase شامل یک میلیون نوع موجودیت و۲۵ هزار نوع رابطه و۱۷ میلیون نمونه برای آموزش و آزمون مدل است. از آنجایی که کار روی داده‌هایی با این تعداد ابعاد بسیار سخت و تقریباً غیر ممکن است باید از روش‌های کاهش ابعاد استفاده کنیم. در این تحقیق سعی می‌کنیم که این داده‌ها را به فضای برداری آورده و مساله را به یک مساله کمینه‌سازی تبدیل کنیم. این داده‌ها را به صورت یک تنسور فرض می‌کنیم که برای هر رابطه یک فضای دوبعدی برداری برای موجودیت‌های مختلف در نظر می‌گیریم، همچنین برای هر رابطه یک بردار بین موجودیت‌های دخیل در آن رابطه خواهیم داشت. سپس تلاش می‌کنیم که از این بردار ها در یک مساله کمینه سازی استفاده کنیم. برای مثال اگه سه‌تایی «head, relation, tail » را داشته باشیم که هرکدام به یک بردار نگاشت شده‌اند انتظار داریم مجموع بردارhead وrelation به بردارtail برسد پس باید مقدار زیر را کمینه کنیم:

$$head + relation - tail => 0$$

برای نمایش این قانون از تخمین مرتبه پایین استفاده می‌کنیم به صورت:

$$f_r(h,t) = ||h + r - t||_2^2$$

در صورتی که تابعf برای سه‌تایی «h, r, t» مقداری نزدیک به صفر داشته باشد این رابطه برقرار است و درغیر این صورت این رابطه برقرار نیست.

در تحقیقاتی که قبلاً روی این مدل داده‌ها انجام شده است فقط از حقایقی که به صورت مستقیم داده شده استفاده شده است. برای مثال اگر در حقایق موجود در داده‌ها داشته باشیم «الف، مادر، ب» و «الف، همسر، ج» فقط از این دو حقیقت استفاده می‌شود و از این حقیقت که «ج، پدر، ب» استفاده نمی‌شود. در این تحقیق تصمیم داریم از قوانین انجمنی بین داده‌ها نیز استفاده کرده و حقایق ضمنی که بین داده‌ها هست را نیز در آموزش مدل استفاده کنیم.

پیشینه تحقیق:

اولین تحقیقاتی که روی این روش استخراج اطلاع از داده‌های چندرابطه صورت گرفت در سال۲۰۱۱ شروع شد و معروف ترین مقالاتی که در این موضوع مطرح شده است عبارتند از **Unstructured Model (UM)** و **Structured Embedding (SE)** و **Single Layer Model (SLM)** و **Semantic Matching Energy (SME)** و **Latent Factor Model (LFM)** و **Neural Tensor Network (NTN)** و **TransE** و **TransH** و **TransR** که در همه موارد فقط از حقایق مستقیم که از پایگاه دانش بدست آمده استفاده شده و از حقایق ضمنی موجود در پایگاه دانش استفاده نشده است. همچنین برای یافتن روابط انجمنی بین داده‌های پایگاه دانش در سال ۲۰۱۳ روش **AMIE** معرفی شده است که قوانین انجمنی را با فرض جهان باز استخراج می‌کند.