Arch

JCWR2020

کنفرانس بین المللی وب پژوهی
ایران، تهران، ۲۸ خرداد ۱۳۹۹

# Expert Finding in Community Question Answering

Mohammad Miri [1], Hamid Beigy [1]

[1]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
smmiri01@gmail.com and beigy@sharif.edu

## Abstract

Community question answering is a forum in which users ask their questions and other existing users respond those questions. As these forums grow, their users, questions and answers increase significantly and new challenges appear. For instance, someone who has asked a question in forum has to wait a lot to another user give him a response. On the other, experts has to spend a lot of time finding questions in their expertise field. Thus, the most important challenge which these forums face is finding appropriate users to answer the questions. One of the most critical issues which expert finding algorithms have is vocabulary gap between skills and words used in posts text. Therefore, in this paper, we propose two translation models to reduce vocabulary gap. The proposed models are evaluated using the StackOverflow dataset and compared with the related algorithms. The experimental results show the superiority of the proposed models.

**Keywords:** Expert Finding, Community Question Answering, Word Embedding, StackOverflow

www.SID.ir

Arch

JCWR2020
کنفرانس بین المللی وب پژوهی
ایران، تهران، ۲۸ خرداد ۱۳۹۹

# جست‌وجوی خبره در سامانه‌های پرسش و پاسخ انجمنی

سید محمد میری[۱]، حمید بیگی[۱]

[۱] دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف، تهران
smmiri01@gmail.com , beigy@sharif.edu

## چکیده

**سامانه‌های پرسش و پاسخ به سامانه‌هایی گویند که کاربران نیاز و پرسش خود را در آن سامانه‌ها ثبت می‌کنند و دیگر کاربران حاضر در سامانه به آن پرسش‌ها پاسخ می‌دهند. با فراگیر شدن این نوع سامانه‌ها تعداد کاربران، پرسش‌ها و پاسخ‌ها در این سامانه‌ها نیز به طور چشمگیری رشد کرد و با این رشد چالش‌های جدیدی برای آن‌ها بوجود آورد. برای مثال ممکن است فردی پرسشی را در سامانه به ثبت برساند و مدت زیادی را انتظار بکشد تا فردی دارای تخصص مربوطه به وی پاسخ دهد، یا از طرفی دیگر ممکن است فردی متخصص نتواند پرسش‌های مربوط به خودش را در سامانه بیابد. پس اصلی‌ترین چالش در عمل یافتن کاربری متخصص برای پاسخ‌دهی به پرسش‌های موجود در زمینه تخصصی خود می‌باشد. یکی از مشکلات مهمی که الگوریتم‌های جست‌وجوی خبره با مواجه‌اند فاصله معنایی میان کلمات به کار رفته داخل متن پرسش و مهارت‌های موجود در سامانه می‌باشد. به همین جهت در این مقاله دو مدل ترجمه ارائه شده است تا این فاصله معنایی را به حداقل برساند. مدل های ارایه شده با استفاده از مجموعه دادگان Stackoverflow ارزیابی شده و با لگوریتم های مشابه مقایسه گردیده است. نتایج آزمایش ها نشان از برتری مدل های پیشنهادی دارد.**

## کلمات کلیدی

**جست‌وجوی خبره، سامانه‌های پرسش و پاسخ انجمنی، تعبیه کلمات،** StackOverflow

## ۱- مقدمه

روش‌های سنتی نظیر استفاده از موتورهای جست‌وجو برای یافتن پاسخ سوالات مشکلات متعددی را به همراه دارند. نخست ممکن است که پاسخ پرسش مورد نظر در منابع موجود موتورهای جست‌وجو موجود نباشد و نتوان با روش های بازیابی اطلاعات آن‌ها را بازیابی کرد. دوم برخی از پرسش ها با ادبیات پیچیده انسان‌ها قابل بیان است و نیاز است که یک انسان به پاسخ‌دهی آن ها اقدام کند.

امروزه سامانه‌های پرسش‌وپاسخ انجمنی[۱] محبوبیت زیادی بین مردم پیدا کرده‌اند. این محبوبیت با رشد چشم‌گیر ترافیک برخی از معروف‌ترین سامانه‌های پرسش و پاسخ مانند stackoverflow[۲]، quora[۳] و Yahoo![۴] در سال‌های اخیر مشهود است. کاربران در این سامانه‌ها می‌توانند با پاسخ‌دهی به سوالات دانش خود را با دیگران به اشتراک بگذارند.

با وجود تمام مزایایی که این نوع از سامانه‌ها در اختیار کاربرانشان قرار می‌دهند با مشکلاتی دست و پنجه نرم می‌کنند. چالش نخست آن‌ها در واقع حجم بالای پرسش‌ها و کاربران سامانه می‌باشد. روزانه شاید ده‌ها هزار پرسش جدید داخل سامانه به ثبت برسد و این علاوه بر آن میلیون‌ها پرسشی است که در گذشته در این سامانه ثبت شده اند. این حجم بالای پرسش‌ها عملا کاربرانی که دنبال پرسش در زمینه تخصص خود هستند را با مشکل بزرگ مواجه می‌سازند[18] . از طرفی دیگر ممکن است فردی پرسشی را در این سامانه‌ها به ثبت برساند و مدت زیادی را منتظر بماند تا کاربری که در آن زمینه به خصوص تخصص دارد به پاسخ‌دهی سوال وی بپردازد. از طرفی دیگر ممکن است فردی دارای تخصص به دنبال پرسش‌های مربوط به تخصص خودش باشد و مدت زیادی را صرف جست‌وجوی پرسش کند[10] .

عمل بازیابی خبره[۵] در واقع به این صورت است که یک پرسمان[۶] از کاربر دریافت می‌شود و از بین تمام کاربران این سامانه آن‌هایی که تخصص لازم برای پاسخ‌دهی به پرسمان مورد نظر را دارند را بازیابی

www.SID.ir

می‌کند[19][24] . بسیاری از مطالعات گذشته نشان می‌دهد هسته کوچکی از کاربران در اصل شروع کننده روند پاسخ‌دهی به پرسش ها هستند[7][20] . به عبارتی دیگر برای پاسخ‌دهی به پرسش‌ها یک جمعیت کوچکی از کاربران همیشه بازیابی می‌شوند. هر پرسش با چند برچسب همراه است که مشخص می‌کند زمینه کلی آن پرسش به چه صورت است و مشکل اصلی که در این مقاله برای حل آن در تلاش هستیم در واقع این است که کلمات داخل متون پرسش با کلماتی که در برچسب‌ها ظاهر می‌شوند فاصله معنایی دارند و نیاز است که یک مدل ترجمه شباهت معنایی میان آن‌ها را پیدا کند. برای مثال فرض کنید که فردی پرسش "تفاوت میان apache و nginx در چیست؟" را در سامانه مطرح کند. این پرسش می‌تواند برچسب‌های nginx، apache و webserver را داشته باشد.

در این مقاله دو مدل برای ترجمه معرفی شده است که هر دوی آن‌ها از تعبیه کلمات استفاده می‌کنند. در مدل‌ها سعی می‌شود که برچسب‌ها و کلمات به فضای برداری از متغیرهای پنهان منتقل شوند و امتیاز خروجی یک تابع از بردارهای آن‌ها باشد. در مدل اول این تابع برابر ضرب داخلی می‌باشد و در مدل دوم این تابع لزوما ضرب نیست و می‌تواند هر تابع دیگری باشد. الگریتم های ارایه شده با استفاده از مجموعه دادگان StackOverflow ارزیابی شده است و بر اساس معیارهای MAP و Precision@k با الگوریتم های مقایسه شده است.

در این مقاله ابتدا به بررسی کارهای پیشین در زمینه جست‌وجوی خبره می‌پردازیم سپس مدل جدید برای بازیابی خبره را معرفی می‌کنیم و نتایج آزمایش‌های انجام شده را ارائه می‌دهیم. در نهایت به چند مورد از کارهای آتی اشاره خواهیم کرد.

## ۲- کارهای پیشین

در زمینه جست‌وجوی خبره تلاش‌های زیادی صورت گرفته و مطالعات زیادی انجام شده است[19][24] . در برخی از پژوهش‌ها سعی شده است تا از مدل موضوعی برای حل مشکل فاصله معنایی بهره گرفته شود[11][12][17] . در این مقالات سعی شده است که موضوعات نهانی که در متن پرسش‌ها و پاسخ‌ها وجود دارد با استفاده از الگوریتم‌های استخراج موضوعی مانند LDA ۰استخراج شوند و سپس مناسب‌ترین کاربران جهت پاسخ‌گویی پرسش مورد نظر بازیابی شوند.

برخی دیگر از پژوهش‌ها از پالایش همکارانه و تجزیه ماتریس برای حل این مشکل استفاده کرده‌اند[3][5][23] . در این پژوهش ها هر کدام از کاربران و پرسش‌ها و پاسخ‌ها به یک فضای برداری از متغیرهای پنهان منتقل می‌شوند و رابطه هر کاربر با هر پرسش با ضرب داخلی بردار متناسب با آن‌ها تخمین زده می‌شود.

در کنار روش‌های موجود روش‌های دیگری نیز از شبکه برای ساخت الگوریتم جست‌وجوی خبره بهره گرفته‌اند[9][16][18][22][25] . در این روش‌ها معمولا پرسش‌ها، پاسخ‌ها و کاربران در شبکه به عنوان گره[۷] در شبکه نمایش داده می‌شود. همچنین روابط میان آن‌ها نیز ابتدا با داده‌های موجود تخمین زده می‌شود و سپس نقش یال را در شبکه بازی می‌کنند. در [21] علاوه بر عناصر یاد شده عبارت‌ها[۸] را نیز ابتدا به فضای برداری برده است و از آن‌ها نیز در شبکه بهره گرفته است.

عناصر موجود در سامانه‌های پرسش و پاسخ همیشه ثابت نیستند و در طول زمان تغییر می‌کنند. برخی از پژوهش‌ها با در نظر گرفتن این موضوع سعی می‌کنند تا رفتار عناصر متغیر با زمان را در این نوع از سامانه‌ها در آینده تخمین بزنند[6][13]. در این پژوهش‌ها تغییرات موضوعی، موضوعات نوظهور، عوض شدن علایق و تخصص کاربران در طول زمان و غیره به عنوان عناصر متغیر با زمان در نظر گرفته شده است.

در کنار روش‌هایی که تا کنون ذکر شده است روش‌های گوناگونی نیز از مدل زبانی برای حل این مسئله استفاده کرده‌اند[1][14][15] . بعضی از روش‌های این دسته به ارائه مدل زبانی برای ترجمه کلمات بکار رفته در داخل متون به مهارت‌های موجود در داخل سامانه پرداخته‌اند0[4] . درمرجع 0 دو مدل برای ترجمه ارائه شده است. مدل اول از اطلاعات متقابل[۹] میان کلمات و برچسب‌های پرسش‌ها بهره می‌برد. در مدل دوم ابتدا کلمات با استفاده از LDA به یک فضای برداری منتقل می‌شود که هر کدام از درایه‌های این بردارها در واقع نشان دهنده ارتباط آن کلمه با یک موضوع است، سپس با استفاده از یک مدل سعی می‌شود برچسب‌ها نیز به یک فضای برداری با همان اندازه منتقل شوند طوری که هر چقدر توزیع موضوعی بین کلمه و برچسب به همدیگر شبیه‌تر باشد ضرب داخلی بردار آن‌ها نیز بالاتر خواهد بود. در مرجع [4] از خوشه‌بندی برای ترجمه کلمات بهره گرفته شده است. ابتدا کلمات را با برداری که از برچسب‌ها ساخته شده است خوشه‌بندی کرده است، سپس کلمات داخل هر خوشه را با بردارهایی که از کلمات ساخته شده اند دوباره خوشه‌بندی[۱۰] می‌کند. در واقع خوشه‌بندی اولیه موضوع کلی را نشان می‌دهد و زیرخوشه دوم نیز زیر موضوع را مشخص می‌کند.

## ۳- مدل پیشنهادی

در این قسمت دو مدل پیشنهاد خواهیم داد که هر دو مبتنی بر تعبیه کلمات[۱۱] خواهد بود. هدف تعبیه بردار در واقع انتقال بردارهای تنک با ابعاد بالا به یک فضای برداری متراکم با ابعاد پایین می‌باشد زیرا استفاده از بردارهای تنک با ابعاد بالا برای یادگیری و آزمون چالش‌های زمانی و محاسباتی به وجود می‌آورد. ابتدا برچسب[۱۲] صحیح هر مدل را مطابق فرمول (۱) بدست می‌آوریم.

$$P_{observed}(q|w) = \frac{tf(q,w)}{tf(w)} \qquad (۱)$$

که در آن $tf(q,w)$ نشان دهنده تعداد دفعاتی است که کلمه $w$ و پرسمان $q$ ظاهر شده‌اند و $tf(w)$ نیز تعداد دفعات تکرار $w$ را نشان می‌دهد.

www.SID.ir

در روش اول ابتدا کلمات و برچسب‌ها[۱۳] با بردارهایی با ابعاد پایین نمایش داده می‌شوند سپس ضرب داخلی این بردارها میزان نزدیکی آن کلمه و آن برچسب‌ها را نشان می‌دهد. ساختار کلی این شبکه را می‌توان در شکل (۱) مشاهده کرد. برای آموزش این شبکه از تابع خطای MSE بهره برده‌ایم و از بهینه‌ساز Adam استفاده کرده ایم.

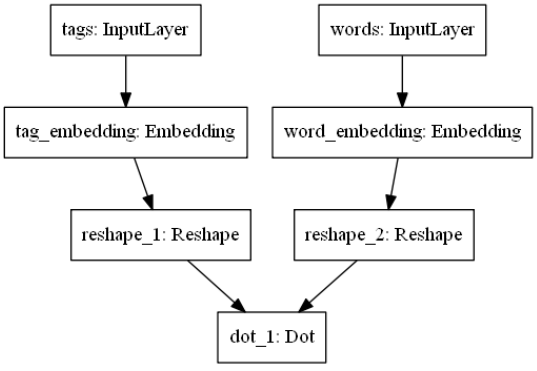
این مدل با مدلی که درمرجع 0 ارائه شد تفاوت‌های اساسی دارد. در مدل مرجع 0 بردار هر کلمه توسط LDA ساخته می‌شود، در این صورت متغیرهای پنهان بردار هر کلمه میزان ارتباط آن کلمه با یک موضوع را نمایش می‌دهد. زمانی که این بردار در داخل مدل قرار می‌گیرد بردار برچسب را نیز به سمتی هدایت می‌کند که شبیه همین بردارهای استخراج شده از LDA باشند. به عبارتی دیگر توزیع موضوعی در کلمات و برچسب‌های شبیه به یکدیگر شبیه هم در بیاید.

در مدل دوم یعنی شبکه عصبی عمیق هم مانند مدل اول همه کلمات و برچسب‌ها به صورت برداری متراکم با ابعاد پایین نمایش داده می‌شوند با این تفاوت که خروجی این تابع دیگر ضرب داخلی نیست. در مدل اول همیشه خروجی مدل برابر ضرب دو بردار در نظر گرفته شده است در حالی که خروجی واقعی می‌تواند یک تابع غیر خطی پیچیده باشد. به همین دلیل در مدل دوم از شبکه عصبی عمیق استفاده می‌کنیم که پارامترهای بسیار بیشتری دارد و می‌تواند توابع بسیار پیچیده‌تری را مدل کند. کلیت ساختار شبکه عمیق را می‌توان در شکل ( ۲) مشاهده کرد.

در مدل دوم دو بردار را صرفا به یکدیگر متصل می‌کنیم و یک بردار بزرگ‌تری از متغیرهای پنهان ایجاد می‌کنیم سپس این بردار بزرگ‌تر به عنوان ورودی به یک شبکه عمیق متصل می‌شود. در این مدل از آنجایی که تابع خروجی لزوما ضرب داخلی دو بردار نیست بردارهای تعبیه می‌توانند به هر اندازه دلخواه باشند و لازم نیست هم اندازه باشند.

بعد از یافتن میزان ارتباط هر کلمه و هر برچسب اکنون می‌توانیم با در اختیار داشتن هر برچسب به عنوان پرسمان ورودی مناسب‌ترین کاربر را جهت پاسخ‌گویی به آن برچسب بازیابی کنیم. در پژوهش‌های مختلف از روش‌های مختلفی برای امتیازدهی کاربران استفاده کرده اند 0[4]. در این پژوهش مدل امتیاز هر کاربر مطابق فرمول (۲) بدست می‌آید.

$$score(e,q) = \sum_{d\in D_e}\sum_{w_i\in D_e} 1_{\{w_i \in translation\ q\}} \qquad (۲)$$




**شکل ۱ :شکل کلیت مدل مبتنی بر تجزیه ماتریس**

## ٤- نتایج آزمایش‌ها

در این قسمت ابتدا معیار ارزیابی را معرفی می‌کنیم سپس به مقایسه مدل‌های ارائه شده با دیگر پژوهش‌ها می‌پردازیم.

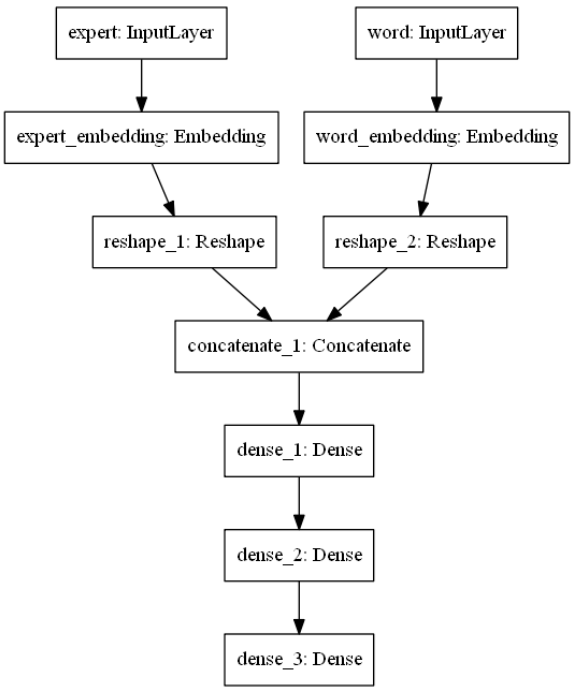
برای ارزیابی مدل از مجموعه دادگان Stackoverflow استفاده کرده ایم. در این مجموعه دادگان پرسش‌ها و پاسخ‌های سال های ۲۰۰۸ تا ۲۰۱۵ وجود دارد. از کل پرسش‌های این مجموعه فقط آن‌هایی که برچسب java دارند مورد استفاده قرار گرفته شده است که تعداد آن‌ها به ۳۱۶،۷۱۶ پرسش و ۳۸۰،۴۲۳،۱ پاسخ می‌رسد.

برای ارزیابی مجموعه پرسمانی که از آن بهره گرفته شده است در واقع ۱۰۰ تا از پرکاربردترین برچسب‌هایی است که با برچسب java ظاهر شده‌اند. گفتنی است که همه کاربران این مجموعه خبره نیستند. برای شناسایی خبرگان از دو معیار استفاده کرده‌ایم. معیار اول این است که کاربران خبره باید حداقل ۱۰ پاسخ برگزیده داشته باشند. معیار دوم آن است که حداقل ۴۰ درصد پاسخ‌هایشان به عنوان پاسخ برگزیده آن سوال انتخاب شده باشد.

دو معیار بسیار رایج برای سنجش خطا در مدل‌های بازیابی اطلاعات وجود دارد که عبارتند از MAP و Precision@k. فرمول کلی Precision@k مطابق فرمول (۳) محاسبه می‌شود.

$$Precision@k = \frac{TP@k}{TP@k+FP@k} \qquad (۳)$$

عبارت $TP$ به مواردی که درست بازیابی شده اند اشاره دارد و عبارت $FP$ مواردی را که به اشتباه بازیابی شده‌اند مشخص می‌کند.




**شکل ۲. کلیت مدل شبکه عصبی عمیق**

www.SID.ir

سپس برای محاسبه MAP ابتدا لازم است که AveragePrecision(AP) را بدست آوریم که مطابق فرمول (۴) بدست می آید.

$$AP = \sum_{n=1}^{N} \frac{rel(n) * Precision(n)}{R} \qquad (۴)$$

که در آن عبارت $rel(n)$ در واقع مشخص می‌کند که مورد بازیابی شده $n$ام با پرسمان مرتبط هست یا خیر و $R$ نیز تعداد کل عبارت‌های درست بازیابی شده است.متغیر MAP نیز در اصل میانگین AP روی همه پرسمان‌ها می‌باشد.

مقدار خطای نهایی مدل تجزیه ماتریس با تغییر ابعاد بردار نهان تغییر می‌کند. همانطور که در شکل (۳) مشاهده می‌کنید با زیاد شدن ابعاد بردار تعبیه مدل پیچیده‌تر می‌شود و نرخ خطا کاهش می‌یابد. وقتی که به ۵۰ می‌رسد بیش برازش[۱۴] رخ می‌دهد و ۴۰ بهترین نتیجه را در خطا به ما می‌دهد.

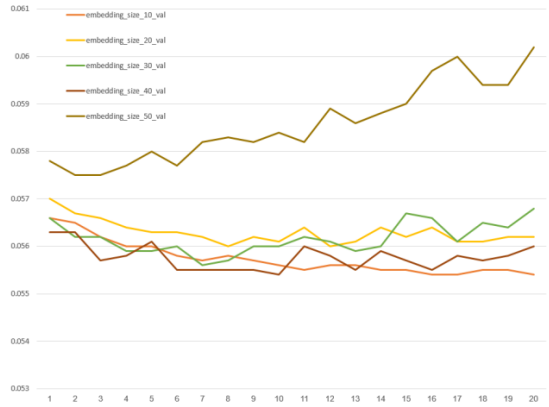
اما تأثیر ابعاد بردار تعبیه در مدل شبکه عمیق ناچیز است. همانطور که در شکل (۴) مشاهده می‌کنید با زیاد شدن ابعاد بردار تعبیه تغییر بخصوصی در خطای مدل نمی‌توان دید.

در شکل‌های (۵) و (۶) بخشی از کلمات استخراج شده برای برچسب‌های مختلف را می‌توان مشاهده کرد.

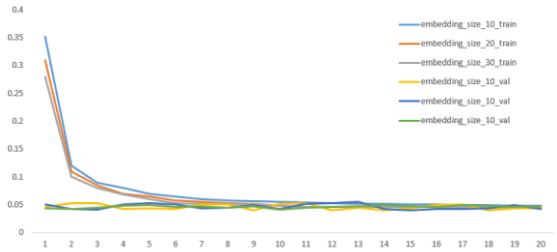
پژوهش مختلفی با همین مجموعه دادگان و همین شرایط خبرگی الگوریتم هایی را ارائه داده‌اند[1][0][4][12] . مقایسه مدل‌های ارائه شده در این پژوهش را با مدل‌های دیگر پژوهش‌ها در جدول (۱) ارایه شده است. همانطور که مشاهده این جدول نشان می دهد هر دو مدل ارائه شده در این پژوهش در معیارهای Precision@1 و MAP از تمام مدل‌های دیگر نتیجه‌ی بهتری را از خود نشان داده‌اند. مدل شبکه عصبی عمیق همچنین در موارد Precision@5 و Precision@10 نیز نتیجه بهتری را رقم زده است.

## ۵- نتیجه

کاربران می‌توانند در سامانه‌های پرسش و پاسخ دانش خود را با دیگران به اشتراک بگذارند. با رشد روز افزون کاربران این نوع از سامانه‌ها چالش‌های جدیدی نیز برای آن‌ها به وجود آمد. مهم‌ترین چالش برای سامانه‌های پرسش و پاسخ در اصل یافتن مناسب‌ترین کاربر جهت پاسخ‌گویی به یک پرسش می‌باشد. در این پژوهش ما با معرفی دو مدل ترجمه سعی در کاهش فاصله معنایی کلمات به کار رفته در متن پرسش‌ها داشتیم. هر دو مدل از تعبیه کلمات بهره برده‌اند. یکی از مدل‌ها از ضرب داخلی بردارها استفاده کرده است ولی در دیگری تابع خروجی لزوما ضرب نیست. همانطور که در نتایج می‌توان دید این دو مدل از همه مدل‌های قبلی در شرایط و مجموعه‌دادگان مشابه نتیجه بهتری از خود نشان دادند.




**شکل ۳. نمودار خطای اعتبار سنجی مدل تجزیه ماتریس برحسب اندازه بردار تعبیه**




**شکل ٤. نمودار خطای مدل شبکه عصبی عمیق برحسب اندازه بردار تعبیه**

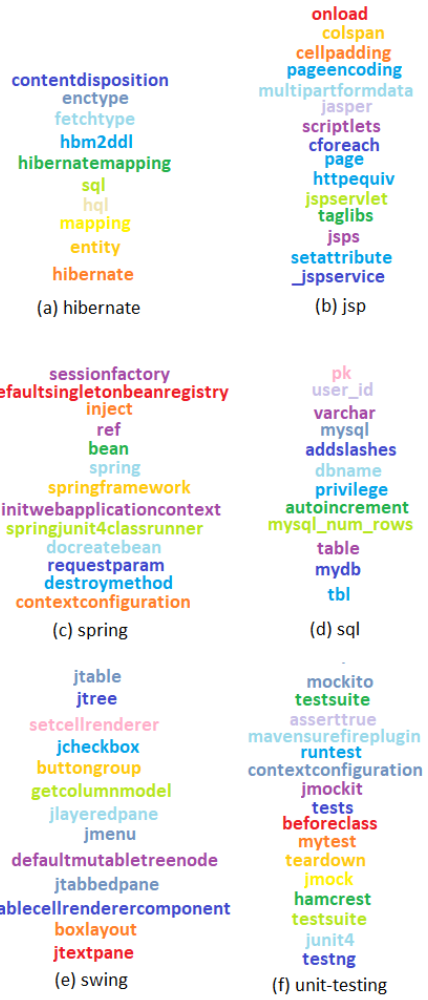
**جدول ۱ : جدول مقایسه عملکر مدل این پژوهش با دیگر پژوهش‌ها**

| روش | P@1 | P@5 | P@10 | MAP |
|---|---|---|---|---|
| روش مبتنی بر مستند[1] | ۵۶ | ۵۰ | ۴۴ | ۳۷.۷ |
| روش مبتنی بر کاربر[1] | ۵۴ | ۴۸.۲ | ۴۲.۵ | ۳۶.۲ |
| روش مدل موضوعی[12] | ۵۵ | ۵۳ | ۴۸.۸ | ۴۳.۴ |
| روش تعبیه کلمات0 | ۶۵ | ۶۲.۶ | ۵۴ | ۴۹.۶ |
| روش اطلاعات متقابل0 | ۶۶ | ۶۰.۴ | ۵۲.۹ | ۴۷.۸ |
| روش خوشه‌بندی[4] | ۶۹.۱ | ۶۱.۵ | ۵۵.۸ | ۵۲.۲ |
| روش تجزیه ماتریس | ۷۲.۷ | ۶۰.۲ | ۴۹.۷ | ۷۳.۳ |
| روش شبکه عمیق | **۸٤.۸** | **٦٦.٤** | **۵۷.۹** | **۷۹.۵** |

## مراجع

[1] Balog, K., Azzopardi, L., & de Rijke, M. (2009). A language modeling framework for expert finding. *Information Processing & Management*, *45*(1), 1-19.

www.SID.ir




شکل ٦ کلمات ترجمه شده توسط مدل شبکه عصبی عمیق

شکل ٥ کلمات ترجمه شده توسط مدل تجزیه ماتریس

[5] Deshpande, M., & Karypis, G. (2004). Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 143-177.

[6] Fang, Y., & Godavarthy, A. (2014, July). Modeling the dynamics of personal expertise. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval* (pp. 1107-1110). ACM.

[7] Fisher, D., Smith, M., & Welser, H. T. (2006, January). You are who you talk to: Detecting roles in usenet newsgroups. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06)* (Vol. 3, pp. 59b-59b). IEEE.

[8] Guo, J., Xu, S., Bao, S., & Yu, Y. (2008, October). Tapping on the potential of q&a community by recommending answer providers. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 921-930). ACM.

[2] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

[3] Cho, J. H., Li, Y., Girju, R., & Zhai, C. (2015, October). Recommending forum posts to designated experts. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 659-666). IEEE.
Dargahi Nobari, A., Sotudeh Gharebagh, S., & Neshati, M. (2017, August). Skill translation models in expert finding. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (pp. 1057-1060). ACM.

[4] Dehghan, M., & Abin, A. A. (2019). Translations Diversification for Expert Finding: A Novel Clustering-based Approach. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *13*(3), 32.

www.SID.ir

Arcl



[9] Jeon, J., Croft, W. B., Lee, J. H., & Park, S. (2006, August). A framework to predict the quality of answers with non-textual features. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 228-235). ACM.

[10] Li, B., & King, I. (2010, October). Routing questions to appropriate answerers in community question answering services. In *Proceedings of the 19th ACM international conference on Information and knowledge management* (pp. 1585-1588). ACM.

[11] Liu, M., Liu, Y., & Yang, Q. (2010, July). Predicting best answerers for new questions in community question answering. In *International Conference on Web-Age Information Management* (pp. 127-138). Springer, Berlin, Heidelberg.

[12] Momtazi, S., & Naumann, F. (2013). Topic modeling for expert finding using latent Dirichlet allocation. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *3*(5), 346-353.

[13] Neshati, M., Fallahnejad, Z., & Beigy, H. (2017). On dynamicity of expert finding in community question answering. *Information Processing & Management*, *53*(5), 1026-1042.

[14] Palangi, H., Smolensky, P., He, X., & Deng, L. (2018, April). Question-answering with grammatically-interpretable representations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[15] Palomera, D., & Figueroa, A. (2017). Leveraging linguistic traits and semi-supervised learning to single out informational content across how-to community question-answering archives. *Information Sciences*, *381*, 20-32.

[16] Pham, T. A. N., Li, X., Cong, G., & Zhang, Z. (2016). A general recommendation model for heterogeneous networks. *IEEE Transactions on Knowledge and Data Engineering*, *28*(12), 3140-3153.

[17] Riahi, F., Zolaktaf, Z., Shafiei, M., & Milios, E. (2012, April). Finding expert users in community question answering. In *Proceedings of the 21st International Conference on World Wide Web* (pp. 791-798). ACM.

[18] Shi, C., Hu, B., Zhao, W. X., & Philip, S. Y. (2018). Heterogeneous information network embedding for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, *31*(2), 357-370.

[19] Wang, X., Huang, C., Yao, L., Benatallah, B., & Dong, M. (2018). A survey on expert recommendation in community question answering. *Journal of Computer Science and Technology*, *33*(4), 625-653.

[20] Welser, H. T., Gleave, E., Fisher, D., & Smith, M. (2007). Visualizing the signatures of social roles in online discussion groups. *Journal of social structure*, *8*(2), 1-32.

[21] Y. Wu and S. Zhao. A novel community answer matching approach based on phrase fusion heterogeneous information networks. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[22] Xiong, Y., Zhu, Y., & Philip, S. Y. (2014). Top-k similarity join in heterogeneous information networks. *IEEE Transactions on Knowledge and Data Engineering*, *27*(6), 1710-1723.

[23] Yang, B., & Manandhar, S. (2014, August). Tag-based expert recommendation in community question answering. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)* (pp. 960-963). IEEE.

[24] Yuan, S., Zhang, Y., Tang, J., Hall, W., & Cabotà, J. B. (2019). Expert finding in community question answering: a review. *Artificial Intelligence Review*, 1-32.

[25] Zhang, J., Ackerman, M. S., & Adamic, L. (2007, May). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web* (pp. 221-230). ACM.

## زیرنویس‌ها

۱ Community Question answering
۲ www.stackoverflow.com
۳ www.quora.com
٤ www.answers.yahoo.com
٥ Expert finding
٦ query
۷ node
۸ phrases
۹ Mutual information
۱۰ clustering
۱۱ Word embedding
۱۲ label
۱۳ tags
۱٤ overfitting

www.SID.ir