# بازیابی و رتبه‌بندی افراد خبره با استفاده از مدل ترجمه مبتنی‌بر خوشه‌بندی

مهدی دهقان[۱]، کارشناسی ارشد؛ احمدعلی آبین[۲]، استادیار

۱- دانشکده مهندسی و علوم کامپیوتر – دانشگاه شهید بهشتی - تهران - ایران – mah.dehghan@mail.sbu.ac.ir
۲- دانشکده مهندسی و علوم کامپیوتر - دانشگاه شهید بهشتی - تهران - ایران – a_abin@sbu.ac.ir

**چکیده:** استخراج دانش از میان داده‌های موجود در وب باتوجه به حجم و تنوع بالای آن به یک چالش در حوزه‌ی بازیابی اطلاعات تبدیل شده‌است. در این میان، مسأله‌ی بازیابی و رتبه‌بندی افراد خبره با هدف بازیابی و رتبه‌بندی افراد خبره در زمینه‌ی موضوع پرس‌وجوی کاربر، به‌عنوان یکی از مهم‌ترین مسائل موجود در این حوزه توجه بسیاری از پژوهشگران را به خود جلب نموده‌است. مهم‌ترین چالش در مسئله‌ی بازیابی افراد خبره تشخیص میزان ارتباط بین کلمات پرس‌وجو و سندهای نوشته‌شده توسط نامزدهای خبرگی است. یک مشکل اساسی در این حوزه فاصله‌ی واژگانی[۱] میان کلمات پرس‌وجو و سندهای نامزدهای خبرگی است. در این مقاله دو مدل ترجمه‌ی[۲] جدید برای مدل‌سازی فاصله‌ی واژگانی ارائه شده‌است. مدل اول یک مدل احتمالاتی مبتنی‌بر خوشه‌بندی و مدل دوم مبتنی‌بر مدل‌سازی موضوعی[۳] است. در هر دو مدل، کلمات پرس‌وجو به مجموعه‌ای از کلمات مرتبط با پرس‌وجو که بیشتر نشان‌دهنده‌ی یک زمینه‌ی خبرگی هستند ترجمه شده‌است. پس از ترجمه‌ی کلمات، از یک مدل ترکیب کننده به‌منظور بازیابی استفاده شده‌است. مدل‌های ارائه‌شده برروی مجموعه‌ی آزمون[۴] Stack Overflow ارزیابی و تحلیل شده‌است. نتایج به‌دست‌آمده بیانگر افزایش میانگین متوسط دقت[۵] روش ارائه‌شده در مقایسه با سایر روش‌های بازیابی افراد خبره است.

**واژه‌های کلیدی:** بازیابی افراد خبره، مدل ترجمه، خوشه‌بندی، مدل‌سازی موضوعی، فاصله‌ی واژگانی، سیستم‌های پاسخ به پرسش.

# Retrieve and Rank the Experts Using a Cluster-based Translation Model

Mahdi Dehghan[1], MSc Student; Ahmad Ali Abin[2], Assistant Professor

1- Faculty of Computer Science and Engineering, University of Shahid Beheshti, Tehran, Iran, Email: mah.dehghan@mail.sbu.ac.ir
2- Faculty of Computer Science and Engineering, University of Shahid Beheshti, Tehran, Iran, Email: a_abin@sbu.ac.ir

**Abstract:** With respect to the increasing volume and variety of information available on the Web, it is very difficult to find the required knowledge through the massive amount of data. Question-answering systems have been created to make easy knowledge accessing through massive amounts of data. The most important factor in the issue of expert finding is the ability to detect the relationship between query words and documents written by the candidate experts. A challenging issue in this area is the vocabulary gap between query words and the documents of the candidate experts. In this paper, two new translation models are proposed to solve the problem of the vocabulary gap. First model, a cluster-based probabilistic model, and another is based on topic modeling. In these models, the query words are translated into a collection of query-related words, which are written in documents written by more candidate experts. Then, using these words and using a simple composite model, we have retrieved the experts. The proposed models are implemented and evaluated on the Stack overflow test set and finally, we have analyzed the outputs. The results indicate an increase in the Mean Average Precision of the proposed method compared with other methods of expert finding.

**Keywords:** Expertise retrieval, translation model, question answering systems, topic modeling, vocabulary gap.

تاریخ ارسال مقاله: ۱۳۹۶/۰۷/۲۶
تاریخ اصلاح مقاله: ۱۳۹۶/۱۲/۱۴
تاریخ پذیرش مقاله: ۱۳۹۷/۰۳/۰۶
نام نویسنده مسئول: احمدعلی آبین
نشانی نویسنده مسئول: ایران - تهران - ولنجک - دانشگاه شهید بهشتی - دانشکده مهندسی و علوم کامپیوتر.

## ۱ -مقدمه

فعالیت گسترده‌ی کاربران در فضای مجازی باعث افزایش روزافزون حجم و تنوع اطلاعات موجود در وب شده‌است و درنتیجه توجه اکثر محققان را به حوزه‌ی بازیابی اطلاعات جلب کرده‌است [۱، ۲]. در این شرایط، وجود سیستمی که بتواند از میان حجم انبوه داده‌ها راه رسیدن به دانش را برای کاربر تسهیل کند به موضوع بسیارمهمی در جامعه‌ی بازیابی اطلاعات تبدیل شده‌است. موتورهای جست‌وجوی کنونی، با بازیابی اسناد مرتبط با پرس‌وجوی کاربر، اطلاعات موردنیاز برای کاربر را فراهم می‌کنند. بااین‌وجود، پیداکردن پاسخ و رسیدن به دانش موردنیاز از میان سندهای زیاد بازیابی‌شده هنوز امری دشوار است. یکی از مسائلی که کاربران در فضای وب با آن روبه‌رو هستند مسأله‌ی بازیابی افراد خبره است. در این مسأله هدف بازیابی و رتبه‌بندی افراد خبره درزمینه‌ی پرس‌وجوی کاربر است. به‌عنوان یک مثال از سیستم‌های بازیابی افراد خبره می‌توان به انتخاب کمیته‌ی علمی مرتبط با موضوع همایش‌ها یا انتساب مقالات علمی به داوران خبره مرتبط با موضوع مقاله اشاره کرد. از دیگر کاربردهای مهم سیستم‌های بازیابی افراد خبره می‌توان به کاربری آنها در شبکه‌های پرسش‌وپاسخ اشاره کرد که در آن معرفی افراد خبره با بهره‌گیری از تاریخچه‌ی پاسخ‌گویی افراد به سوالات صورت می‌گیرد. به‌عنوان‌مثال، شبکه‌ی Stack Overflow یکی از شبکه‌های پرسش و پاسخ فعال در بستر وب است که در آن کاربران می‌توانند سوالات خود را مطرح کنند یا به سوالات دیگران پاسخ دهند. کاربران می‌توانند به سوال‌های مطرح‌شده یا پاسخ‌های داده‌شده رأی داده و اهمیت و کیفیت سوال‌ها و جواب‌ها را مشخص نمایند. هر کاربری که سوالی را مطرح می‌کند، باید به آن سوال یک یا چند برچسب بزند، این برچسب‌ها مهارت‌های لازم برای پاسخ‌گویی به سوال مطرح شده را مشخص می‌کند. شرکت‌هایی که به‌دنبال استخدام افراد خبره در زمینه‌های مختلف هستند می‌توانند با استفاده از پاسخ‌هایی که نامزدهای خبرگی دهده‌اند افراد خبره را شناسایی کرده و به آن‌ها شغل پیشنهاد دهند. فرآیند بازیابی افراد خبره یک فرآیند مبتنی‌بر یادآوری[۶] است چون صاحبان مشاغل ترجیح می‌دهند تاجایی‌که ممکن است افراد خبره‌ی بیشتری را شناسایی کنند و سپس طی فرآیندهایی مانند مصاحبه یا آزمون افراد موردنیاز خود را انتخاب کنند.

در شکل (۱) یک صفحه‌ی نوعی از شبکه‌ی پرسش و پاسخ Stack Overflow به‌همراه اجزای مهم اطلاعات آن نشان داده شده‌است. سه جزء اطلاعاتی مهم در این صفحه عبارتند از: ۱) یک سوال که توسط یک کاربر در این شبکه مطرح می شود، ۲) یک یا چند برچسب که توسط کاربر مطرح‌کننده سوال به آن سوال منتسب می‌شود و نشان‌دهنده‌ی مجموعه‌ی مهارت‌های لازم برای پاسخ‌گویی به آن سوال است و ۳) مجموعه‌ی پاسخ‌ها که توسط دیگر کاربران به آن سوال داده شده‌است.


**شکل ۱: یک صفحه نوعی از شبکه پرسش‌وپاسخ Stack Overflow**

دو مدل زبانی پایه که می‌توان از آن‌ها برای بازیابی و رتبه‌بندی افراد خبره در شبکه‌ی Stack Overflow استفاده کرد مدل‌هایی هستند که بلوگ و همکارانش در [۳] ارائه کرده‌اند. در این مدل‌ها می‌توان از برچسب‌های سوال به‌عنوان پرس‌وجو و از پاسخ‌های داده شده به عنوان اسناد استفاده کرد. برای نمونه برچسب Mysql می‌تواند به عنوان یک پرس‌وجو درنظر گرفته‌شود. مهم‌ترین ایراد این مدل‌ها فاصله‌ی واژگانی است برای مثال شخصی که در Mysql خبره است شاید به‌ندرت در پاسخ‌های خود از کلمه‌ی Mysql استفاده کند. روش‌های زیادی برای رفع این مشکل ارائه شده‌اند که از بین آن‌ها می‌توان به روش‌های ارائه‌شده در [۴، ۵، ۶، ۷] اشاره کرد.

مهم‌ترین چالش مطرح در مسئله بازیابی و رتبه‌بندی افراد خبره، فاصله‌ی واژگانی کلمات پرس‌وجو و محتوای سندها (پاسخ افراد در سامانه‌های پاسخ به پرسش) است. در این مقاله هر کلمه‌ی پرس‌وجو به مجموعه‌ای از کلمات مرتبط ترجمه می‌شود و سپس با استفاده از این کلمات، عملیات بازیابی افراد خبره انجام می‌گیرد. با این کار سعی بر آن شده‌است که تا حد ممکن مسأله‌ی فاصله‌ی واژگانی در شبکه‌ی پرسش و پاسخ لحاظ شود.

### ۱ ۱ -مروری بر کارهای گذشته

مهم‌ترین وظیفه‌ی سیستم‌های بازیابی افراد خبره، بازیابی و رتبه‌بندی افراد خبره در زمینه‌ی پرس‌وجوی کاربر می‌باشد. فرآیند بازیابی افراد خبره یک فرآیند مبتنی‌بر یادآوری است به‌عبارتی‌دیگر در این سیستم‌ها باید تا جایی که ممکن است افراد خبره شناسایی شوند. چالش اساسی در فرآیند بازیابی افراد خبره نحوه‌ی مدل کردن ارتباط بین کلمات پرس‌وجو و نامزدهای خبرگی و اندازه‌گیری میزان قدرت این ارتباط است[۸]. ازجمله مدل‌های موفق در سیستم بازیابی افراد خبره می‌توان به مدل‌های مولد احتمالاتی[۳، ۱۱]، مبتنی‌بر رأی‌گیری

و مبتنی‌بر گراف اشاره کرد که در ادامه به هر یک با جزییات بیش‌تری پرداخته می‌شود. در جدول (۱) نشانه‌ها و علائم به‌کاررفته در مقاله به‌منظور درک بهتر روابط و توضیحات مقاله آورده شده‌است.

**جدول ۱: جدول علائم و نشانه‌های به کار رفته در مقاله**

| نماد | توضیحات |
|---|---|
| d | سند |
| q | پرس‌وجو |
| e | نامزد خبرگی |
| t | برچسب یا زمینه‌ی خبرگی |
| w | کلمه |

**مدل‌های مولد احتمالاتی:** در این مدل‌ها[۷] نامزدهای خبرگی براساس مقدار احتمال $P(e|q)$ رتبه‌بندی می‌شوند. این احتمال نشان‌دهنده‌ی این است که کلمات پرس‌وجوی q چه اندازه باعث تولید نامزد خبرگی e می‌شوند. برای محاسبه‌ی این احتمال دو روش ارائه شده‌است. در [۱۱] مقدار $P(e|q)$ به‌صورت مستقیم و با استفاده از یک مدل زبانی دوبخشی محاسبه می‌شود. تفسیر این احتمال این است که پرس‌وجوی q با چه احتمالی باعث تولید نامزد خبرگی e می‌شود. به همین دلیل به این مدل، مدل مولد نامزد[۸] می‌گویند. در بخش اول که به آن مدل ارتباطی نیز گفته‌می‌شود میزان مرتبط‌بودن پرس‌وجو و یک سند با استفاده از مدل زبانی آن سند محاسبه می‌شود. در بخش دوم که به آن مدل هم وقوعی نیز گفته‌می‌شود میزان مرتبط‌بودن یک نامزد به یک سند با دادن پرس‌جو محاسبه می‌شود. از حاصل ضرب مقدار خروجی این دو بخش ارزش هر سند در خبره بودن یک شخص محاسبه می‌شود. درنتیجه برای هر نامزد خبرگی مجموع این ارزش‌ها به‌عنوان میزان خبرگی آن نامزد در موضوع پرس‌وجوی کاربر درنظر گرفته‌می‌شود.

بلوگ و همکارانش در [۳] برای محاسبه‌ی احتمال p(e|q) از قاعده‌ی بیز استفاده کرده‌اند و دو مدل با نام‌های *مدل مبتنی بر نمایه* و *مدل مبتنی بر سند* ارائه‌کرده‌اند. در مدل مبتنی‌بر نمایه برای هر نامزد خبرگی براساس سندهایش یک مدل زبانی ایجاد می‌شود. در واقع این مدل‌های زبانی نشان‌دهنده‌ی نمایه‌ی هر خبره است. در این مدل فرض می‌شود کلمات پرس‌وجو از یکدیگر مستقل هستند در نتیجه احتمال رخ‌دادن هر یک از کلمات پرس‌وجو در نمایه‌ها محاسبه و در هم ضرب می‌شوند. درنهایت مقدار این حاصل‌ضرب به عنوان میزان خبرگی هر نامزد درنظر گرفته‌می‌شود. در مدل مبتنی‌بر سند ابتدا اسناد مرتبط با پرس‌وجوی کاربر بازیابی‌شده سپس ارزش هر سند بازیابی‌شده براساس احتمال رخ‌دادن پرس‌وجو در آنها محاسبه می‌شود. درنهایت هر نامزد متناسب با سهمی که در یک سند بازیابی‌شده دارد از ارزش آن سند امتیاز کسب می‌کند. مجموع امتیازات کسب‌شده برای هر نامزد میزان خبرگی‌اش را در موضوع پرس‌وجوی کاربر نشان می‌دهد.

**مدل‌های مبتنی‌بر رأی‌گیری:** مدل‌های مبتنی‌بر رأی‌گیری[۹] از تکنیک‌های ادغام داده‌ها الهام گرفته‌اند. مفهوم ادغام داده‌ها در بازیابی افراد خبره را می‌توان این‌گونه درنظر گرفت که سندهایی که براساس پرس‌وجوی کاربر رتبه‌بندی می‌شوند به نامزدهای خبرگی که‌درآن ذکر شده‌اند یا سند متعلق به آن است رأی می‌دهند [۸]. در [۹] ابتدا اسناد با استفاده از پرس‌وجوی کاربر بازیابی می‌شوند سپس هر سند بازیابی شده به نویسنده‌هایش ۱ امتیاز می‌دهد. درنهایت نویسنده‌ها بر اساس مجموع امتیازات کسب‌شده رتبه‌بندی می‌شوند. در [۱۰] ابتدا اسناد مرتبط با پرس‌وجوی کاربر بازیابی می‌شوند سپس هر سند بازیابی‌شده به‌اندازه‌ی معکوس رتبه‌اش به نویسنده‌هایش امتیاز می‌دهد. در نهایت نویسنده‌ها براساس مجموع امتیازات کسب‌شده رتبه‌بندی می‌شوند.

**مدل‌های مبتنی‌بر گراف:** روش‌هایی که تاکنون به آن‌ها اشاره شد شواهد خبرگی را از تحلیل محتوای متنی سندها استخراج می‌کردند. می‌توان برای بهبود در سیستم‌های بازیابی افراد خبره از روش‌های مبتنی‌بر گراف استفاده کرد. در روش‌های مبتنی‌بر گراف می‌توان نامزدهای خبرگی و سندها را به‌عنوان رئوس گراف خبرگی درنظر گرفت و گراف خبرگی را ترسیم نمود. در [۱۲، ۱۳، ۱۴] ادعا شده است که انتقال اطلاعات ناشی از ارتباطات شخصی و غیر رسمی است. در گراف خبرگی نیز می‌توان از این ادعا برای بازیابی افراد خبره استفاده کرد. به‌طورکلی روش‌های مبتنی‌بر گراف به دو دسته‌ی *وابسته به پرس‌وجو* و *مستقل از پرس‌وجو* دسته‌بندی می‌شوند.

علاوه‌بر مدل‌های فوق، مدل‌های دیگری نیز در حوزه بازیابی و رتبه بندی افراد خبره ارائه‌شده‌اند که به‌صورت پراکنده یا ترکیبی کار کرده‌اند. به‌دلیل نزدیکی زیادی که بازیابی افراد خبره با بازیابی سند و سایر مفاهیم بازیابی موجود در بازیابی اطلاعات دارد، روش‌های کلاسیک بازیابی اطلاعات را نیز می‌توان در بازیابی افراد خبره به کار برد. از جمله روش‌های به‌کاربرده‌شده می‌توان به مدل‌سازی موضوعی[۱۰] [۱۵] و مدل فضای برداری[۱۱] [۱۶] اشاره کرد. یکی از روش‌های مدل‌سازی موضوعی که در بازیابی افراد خبره استفاده‌شده‌است روش تخصیص پنهان دریکله[۱۲] [۱۵] است. در این روش، ابتدا ماتریس سند-کلمه که خانه‌های آن نشان‌دهنده‌ی تعداد رخداد یک کلمه در یک سند است را تشکیل می‌دهیم و سپس آن را به دو ماتریس کلمه-موضوع و موضوع-سند تجزیه می‌کنیم. روش ارائه‌شده در [۶] از روش تخصیص پنهان دریکله برای بازیابی افراد خبره استفاده می‌کند. در این روش در قدم اول با استفاده از روش تخصیص پنهان دریکله موضوع‌های اصلی هر سند استخراج می‌شود سپس از این موضوع‌ها به‌عنوان یک واسطه تخمین احتمال خبره بودن یک نامزد خبرگی در زمینه‌ی پرس‌وجو استفاده می‌شود. در [۱۶] از یک شبکه‌ی عصبی سه لایه که یک لایه‌ی مخفی دارد، جهت بازیابی و رتبه‌بندی افراد خبره استفاده شده‌است. در لایه‌ی خروجی این شبکه به اندازه‌ی تعداد نامزدهای خبرگی در مجموعه‌ی آزمون نورون وجود دارد. در لایه‌ی ورودی این شبکه هر یک از کلمات پرس‌وجو به‌صورت یک بردار که فقط یک عنصر یک دارد و بقیه صفر هستند (one-hot) و به صورت

مستقل از هم به شبکه اعمال می‌شوند سپس از طریق مجموعه وزن‌های بین نورون‌های لایه‌ی ورودی و لایه‌ی مخفی این بردارها به یک بردار در فضای جدیدی نگاشت می‌شوند طوری‌که کلمات شبیه به هم فاصله‌ی کمی باهم داشته باشند. سپس بااستفاده از مجموعه‌ی وزن‌های بین نورون‌های لایه‌ی مخفی و خروجی، که در واقع میزان تأثیر هر بعد از فضای جدید در خبره بودن هر نامزد را نشان می‌دهد، بردارهای فضای جدید به یک بردار به اندازه‌ی تعداد نامزدهای خبرگی نگاشت می‌شوند به‌طوری‌که هر عنصر از بردار خروجی نشان‌دهنده‌ی میزان خبرگی نامزد متناظرش در کلمه‌ی پرس‌وجوی واردشده به شبکه است. پس از اعمال همه‌ی کلمات پرس‌وجو به شبکه، میزان خبرگی محاسبه‌شده برای هریک از نامزدها در هم ضرب‌شده و نامزدها براساس مقدار خروجی این حاصل ضرب رتبه‌بندی می‌شوند. در [۱۷] روشی برای بازیابی و رتبه‌بندی افراد خبره در سیستم پرسش‌وپاسخ کوئرا ارائه شده‌است. آن‌ها با تحلیل رفتار افراد خبره و غیرخبره دریافتند تفاوت قابل توجهی بین آن‌ها وجود دارد. از این رو با استخراج ویژگی‌هایی که بتواند به خوبی اشخاص خبره و غیر خبره را از هم جدا کند، مسأله‌ی بازیابی و رتبه‌بندی افراد خبره را به یک مسئله‌ی دسته‌بندی تبدیل کردند.

یکی از مشکلات مطرح در مدل‌های مولد احتمالاتی فاصله‌ی واژگانی بین کلمات پرس‌وجو و سندهای نامزدهای خبرگی است. در این روش‌ها از احتمال رخداد کلمات پرس‌وجو برای بازیابی و رتبه‌بندی افراد خبره استفاده می‌شود. ممکن است شخصی در سندهایش به جای استفاده از خود کلمات پرس‌وجو از کلماتی استفاده کند که بیشترین ارتباط معنایی را با کلمات پرس‌وجو داشته باشد، مدل‌های مولد احتمالاتی ارائه‌شده در [۳] به این شخص امتیاز کمی می‌دهند زیرا آن‌ها فقط از احتمال رخ دادن خود کلمات پرس‌وجو برای بازیابی و رتبه‌بندی افراد خبره استفاده می‌کنند. از آنجاکه در مدل‌های ترجمه، یک کلمه با احتمال مشخص به دیگر کلمات ترجمه می‌شود، انتظار می‌رود که با استفاده از این مدل‌ها بتوان مشکل فاصله واژگانی بین کلمات پرس‌وجو و سندهای نامزدهای خبرگی را حل کرد [۴، ۱۹، ۲۰، ۲۱]. در این مدل‌ها ابتدا کلمات پرس‌وجو به مجموعه‌ای از کلمات مرتبط به پرس‌وجو ترجمه شده سپس بازیابی افراد خبره با استفاده از این کلمات انجام می‌شود. در [۴] یک مدل ترجمه‌ی احتمالاتی ارائه شده‌است که مبتنی‌بر اطلاعات متقابل[۱۳] بین کلمات است. در [۱۹] یک مدل ترجمه‌ی مبتنی‌بر اطلاعات متقابل کلمات و درهم‌سازی کلمه[۱۴] برای رفع این مشکل ارائه شده‌است. مدل‌های ترجمه در حوزه‌های زیادی از بازیابی اطلاعات ازجمله بازیابی اطلاعات متقابل زبانی[۱۵] [۲۲، ۲۳، ۲۴] و بازیابی جمله[۱۶] [۲۵] کاربرد دارند.

### ۱ ۲ -خوآوری مقاله

در این مقاله دو مدل ترجمه‌ی زبانی کارآمد برای رفع مشکل فاصله‌ی واژگانی بین کلمات پرس‌وجو و سندهای نوشته‌شده توسط نامزدهای خبرگی پیشنهاد شده‌است. مدل اول، یک مدل احتمالاتی مبتنی‌بر خوشه‌بندی[۱۷] و مدل دوم، مبتنی‌بر مدل‌سازی موضوعی است. در مدل اول با استفاده از دو مفهوم درجه‌ی تعلق کلمات به خوشه‌ها و میزان شباهت خوشه‌ها به برچسب‌ها در یک چارچوب احتمالاتی، سعی شده‌است مشکل فاصله‌ی واژگانی تا حد ممکن رفع‌شود. در مدل دوم، مهم‌ترین موضوعاتی که هر پرس‌وجو یا برچسب به آن‌ها مرتبط است با استفاده از مدل‌سازی موضوعی مشخص شده و درنهایت با استفاده از کلمات کلیدی هر موضوع، ترجمه‌ی یک برچسب را مشخص کرده و مسئله‌ی فاصله‌ی واژگانی تا حد ممکن برطرف شده‌است.

### ۱ ۳ -ساختار مقاله

روش پیشنهادی در بخش ۲ به‌طور کامل بیان شده‌است. بخش ۳ پیاده‌سازی روش پیشنهادی را به‌همراه مجموعه‌ی آزمون[۱۸] و پارامترهای مسئله بیان می‌کند. در بخش ۴ خروجی‌های روش پیشنهادی ارائه و تحلیل شده‌است و درنهایت مقاله در بخش ۵ با جمع‌بندی و نتیجه‌گیری پایان یافته‌است.

## ۲ -روش پیشنهادی

همان‌گونه‌که در بخش ۱ عنوان شد، ساختار شبکه Stackoverflow بدین‌گونه طراحی شده‌است که هر سوال یک یا چند برچسب مخصوص به خود دارد که طراح آن سوال با توجه به ذات پرسش مشخص و به سوال منتسب می‌کند. می‌توان از این برچسب‌ها به‌عنوان کلمات پرس‌وجو استفاده کرد. بدین معنی که کاربر در حین استفاده از مدل پیشنهادی برای خبره‌یابی یک (یا چند) برچسب از این برچسب‌ها را به‌عنوان کلمه (یا کلمات) پرس‌وجو انتخاب کرده و با تکنیک‌های ترجمه‌ای که در این مقاله ارائه شده خبره‌ها را بازیابی می‌کند. در این مقاله از ۱۰۰ برچسب پررخداد با برچسب java به‌عنوان پرس‌وجو استفاده شده‌است. وظیفه‌ی مدل‌های ترجمه، تبدیل یا ترجمه‌ی کلمات پرس‌وجوی کاربر (که دراینجا از بین برچسب‌ها انتخاب می شوند) به مجموعه‌ای از کلمات است که با احتمال بیشتری حاوی اطلاعات خبرگی هستند. در ادامه، مدل‌های پیشنهادی با جزییت توصیف شده‌است.

### ۲ ۱ -مدل یک: مدل احتمالاتی مبتنی‌بر خوشه‌بندی

در این مدل، از تکنیک خوشه‌بندی برای ارائه یک مدل ترجمه برای پرس‌وجوی کاربر استفاده شده‌است. مدل پیشنهادی مبتنی‌بر شش گام است که به شرح زیر می باشد:

**گام ۱: استخراج کلمات مهم با استفاده از مدل‌سازی موضوعی**

گام اول درواقع نوعی پیش‌پردازش برروی کل سندها است که به‌منظور استخراج مهم‌ترین کلمات موجود در اسناد انجام شده‌است. همان‌گونه‌که بیان شد، هدف اصلی مدل‌های پیشنهادی در این مقاله، ترجمه‌ی کلمات پرس‌وجو به مجموعه‌ای از کلمات است که بیشترین ارتباط را با پرس‌وجوی کاربر دارند. به‌منظور ازبین‌بردن کلمات

بی‌ارزش و بی‌ربط مانند حروف اضافه، قیود و صفات که در پاسخ‌ها وجود دارند اولین گام، انجام مدل‌سازی موضوعی است.

در این گام برای انجام مدل‌سازی موضوعی، از ابزار MALLET استفاده شده‌است. در ابزار MALLET فایل ورودی باید در قالب خاصی که توسط توسعه‌دهندگان آن تعریف شده‌است تهیه شود. قالب ورودی که در این مدل از آن استفاده شده‌است به‌این‌صورت است که هر سند (همان پاسخ‌داده‌شده به سوال در Stack Overflow) به‌همراه شناسه‌ی آن سند در یک سطر از فایل ورودی قرار داده شده و سپس این فایل را در اختیار ابزار MALLET قرار داده و با استفاده از روش تخصیص پنهان دریکله، سندها در ۱۰۰ موضوع دسته‌بندی می‌شوند. پس از اتمام عملیات مدل‌سازی موضوعی، ۵۰۰ کلمه‌ی نخست هر موضوع که درواقع بیشترین ارتباط با آن موضوع دارند به عنوان کلمات مهم در بازیابی افراد خبره درنظر گرفته می‌شوند.

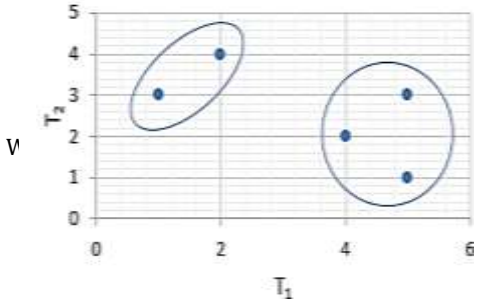
**گام ۲: تشکیل ماتریس کلمه-برچسب**

در این گام، ماتریس کلمه-برچسب $WTF \equiv \left[ wtf_{ij} \right]_{n_w * n_t}$ که هر خانه‌ی $wtf_{ij}$ از این ماتریس نشان‌دهنده‌ی تعداد رخداد کلمه‌ی i ام با برچسب j ام است تشکیل می‌شود.

$$wtf_{ij} = tf(\mathrm{w}_i, \mathrm{t}_j) \qquad (۱)$$

**گام ۳: خوشه‌بندی کلمات**

در این گام، نرمال‌شده‌ی بردار کلمات موجود در فضای ۱۰۰ بعدی برچسب‌ها (نرمال‌شده‌ی هر سطر از ماتریس کلمه-برچسب) خوشه‌بندی می شوند تا با استخراج خوشه‌ها بتوان کلماتی که در یک طیف معنایی قرار می‌گیرند را شناسایی نمود. بدیهی است که در این مدل، هر کلمه به‌عنوان یک بردار ۱۰۰ بعدی در فضای برچسب‌ها می‌باشد. خوشه‌بندی کلمات در این فضا به‌معنی قراردادن کلمات مشابه یا با طیف معنایی خاص در یک خوشه است. از اطلاعات این خوشه‌ها در گام بعد به‌منظور لحاظ‌کردن فاصله‌ی واژگانی استفاده شده‌است. به‌این‌صورت که به‌جای استفاده از خود کلمه‌ی پرس‌وجو دریافتی، از میزان تعلق کلمه به یک خوشه و میزان تعلق آن خوشه به برچسب استفاده می‌شود. این امر سبب می‌شود تا ناکارآمدی یک کلمه در بازیابی برچسب مناسب به‌کمک کلمات هم‌خوشه آن تعدیل شود. برای مثال فرض کنید گام اول (بخش ۱-۲) با پردازش چند سند، ۵ کلمه را به‌عنوان مهم‌ترین کلمات خروجی داده‌است و سندهای ما در کل شامل ۲ برچسب می‌باشند. حال فرض کنید که گام ۲، ماتریس کلمه-برچسب به‌صورت شکل (۲) را نتیجه داده است. خوشه بندی سطرهای این ماتریس که بیان‌گر کلمات در فضای دوبعدی برچسب‌ها هستند به‌صورت شکل (۲) است که به‌گونه‌ای بیان‌گر قرارگرفتن کلمات مشابه در یک خوشه است.




**شکل ۲: مثالی از خوشه‌بندی کلمات در فضای برچسب‌ها**

**گام ۴: محاسبه‌ی احتمال ترجمه‌ی پرس‌وجو به کلمه**

در این گام کلمه‌ی پرس‌وجوی موردنظر کاربر (که از مجموعه برچسب‌های پرتکرار با جاوا انتخاب شده‌اند) باید به مجموعه‌ای از کلمات مرتبط ترجمه شود یا به‌عبارتی‌دیگر مدل قادر به بسط کارآمد پرس‌وجو باشد. برای یافتن ترجمه‌ی هر برچسب یا پرس‌وجو باید مقدار احتمال $P(w|t)$ محاسبه شود. درواقع این احتمال نشان‌دهنده‌ی احتمال ترجمه‌ی یک برچسب t به یک کلمه‌ی w را نشان می‌دهد. محاسبه‌ی این احتمال با استفاده از روابط (۲) و (۳) انجام می‌گیرد.

$$P(\mathrm{w} \mid \mathrm{t}) = \frac{P(\mathrm{t} \mid \mathrm{w}) * \mathrm{P}(\mathrm{w})}{P(\mathrm{t})} = \frac{P(\mathrm{t} \mid \mathrm{w}) * \mathrm{P}(\mathrm{w})}{\sum_{i=1}^{n_w} P(t \mid w_i) * P(w_i)} \qquad (۲)$$

در رابطه‌ی (۲)، P(w) احتمال پیشین رخدادن کلمه‌ی w است که برای همه‌ی کلمات ثابت درنظر گرفته شده‌است، P(t) احتمال پیشین رخدادن برچسب t است ازآن‌جاکه در رتبه‌بندی نهایی تأثیری ندارد نیازی به محاسبه‌ی آن نیست. P(t|w) نیز احتمال تولید برچسب t توسط کلمه‌ی w است که برای محاسبه‌ی آن از میزان مرتبط‌بودن خوشه به برچسب (P(t|c)) و میزان متعلق‌بودن کلمه به خوشه (P(c|w)) به‌صورت زیر استفاده می‌شود:

$$\begin{aligned} P(\mathrm{t} \mid \mathrm{w}) &= \sum_{i=1}^{n_c} P(\mathrm{t}, \mathrm{c}_i \mid \mathrm{w}) \\ &= \sum_{i=1}^{n_c} P(\mathrm{t} \mid \mathrm{c}_i, \mathrm{w}) * P(\mathrm{c}_i \mid w) \\ &= \sum_{i=1}^{n_c} P(t \mid c_i) * P(c_i \mid w) \\ P(\mathrm{w}) &= \frac{1}{n_w} \text{ for all of words} \end{aligned} \qquad (۳)$$

که درآن c بیانگر خوشه، $n_c$ بیانگرتعداد خوشه‌ها است که در الگوریتم خوشه‌بندی انتخاب شده‌است، w بیانگر کلمه، $n_w$ بیانگر تعداد کل کلمه‌ها، که درواقع مجموع ۵۰۰ کلمه‌ی اول هر موضوع در مدل‌سازی موضوعی است و t بیانگر برچسب یا همان زمینه‌های خبرگی است.

برای محاسبه‌ی $P(t|c_i)$ که نشان می‌دهد خوشه‌ی $c_i$ با چه احتمال باعث تولید برچسب t می‌شود از رابطه‌ی زیر استفاده می‌شود:

$$P(t \mid c_i) = \frac{\frac{1}{d(t, c_i)}}{\sum_{j=1}^{n_c} \frac{1}{d(t, c_j)}} \quad (۴)$$

در رابطه‌ی (۴)، $d(t,c_i)$ نشان‌دهنده‌ی میزان فاصله‌ی کسینوسی بین مرکز خوشه‌ی $c_i$ تا بردار یکه‌ی متناظر با برچسب t است.

برای محاسبه‌ی $P(c_i|w)$ که نشان‌دهنده‌ی احتمال تولید خوشه‌ی $c_i$ توسط کلمه‌ی w است از رابطه‌ی زیر استفاده می‌شود:

$$P(\mathrm{c}_i \mid \mathrm{w}) = \frac{\frac{1}{d(\mathrm{c}_i, \mathrm{w})}}{\sum_{j=1}^{n_c} \frac{1}{d(\mathrm{c}_j, \mathrm{w})}} \quad (۵)$$

در رابطه‌ی (۵)، $d(c_i,w)$ نشان‌دهنده‌ی میزان فاصله‌ی کسینوسی بین بردار کلمه‌ی w و مرکز خوشه‌ی $c_i$ است.

**گام ۵: انتخاب بهترین ترجمه‌ها برای هر پرس‌وجو**

در این گام مقدار $P(w|t)$ برای هر یک از ۱۰۰ برچسب به‌صورت نزولی مرتب می‌شود و برای نمونه ۲۰ کلمه‌ی نخست که درواقع بیشترین ارتباط با آن برچسب را دارند به‌عنوان ترجمه‌های یک برچسب انتخاب می‌شود.

**گام ۶: بازیابی و رتبه‌بندی افراد خبره**

پس از مشخص‌شدن ترجمه‌ی پرس‌وجو، باید با استفاده از آن کلمات افراد خبره را شناسایی شوند. در بازیابی افراد خبره، سندها یا جواب‌هایی که کلمات ترجمه در آن‌ها به‌کاربرده‌شده را بازیابی کرده و به نویسنده‌ی آن جواب یک امتیاز داده می‌شود. در نهایت با استفاده از رابطه (۴) مجموع امتیازات هر نویسنده را محاسبه کرده و نویسنده‌ها براساس امتیازات کسب‌شده به صورت نزولی مرتب می‌شوند.

$$Score(e, \mathrm{t}) = \sum_{\{d:e\in d, d\in R(Translation(t))\}} 1 \quad (۶)$$

در رابطه‌ی (۶)، $R(Translation(t))$ (سوال الف ۵) بیانگر مجموعه‌ی سندهایی است که حداقل یکی از کلمات ترجمه‌ی پرس‌وجو در آن به‌کار رفته باشد.

### ۲ ۴ -مدل دو: مدل مبتنی بر مدل‌سازی موضوعی

در این مدل ابتدا برچسب‌ها با استفاده از مدل‌سازی موضوعی در ۱۰۰ موضوع دسته‌بندی شده و سپس مرتبط‌ترین کلمات هر موضوع برای ترجمه‌ی برچسب‌ها انتخاب می‌شود. درنهایت از این ترجمه‌ها در عملیات بازیابی و رتبه‌بندی افراد خبره استفاده می‌شود. این مدل در چهار گام عمل بازیابی و رتبه‌بندی افراد خبره را انجام می‌دهد که به شرح زیر است:

**گام ۱: استخراج رابطه‌ی برچسب‌ها و موضوعات**

در این گام از مدل‌سازی موضوعی به‌منظور استخراج رابطه‌ی برچسب‌ها با موضوعات استفاده می‌شود. برای انجام مدل‌سازی موضوعی از ابزار MALLET استفاده شده‌است. قالب فایل ورودی ابزار MALLET که در این مدل از آن استفاده شده به این صورت است که مجموعه‌ی سندها یا همان پاسخ‌هایی که دارای برچسب خاص هستند را با یک شناسه‌ی خاص در یک سطر از فایل ورودی قرار داده و سپس این فایل را توسط ابزار MALLET و با استفاده از روش تخصیص پنهان دریکله پردازش کرده و برچسب‌ها در ۱۰۰ موضوع دسته‌بندی می‌شوند. قالب ورودی ابزار MALLET در این مدل با مدل قبل که در آن هر سند به همراه شناسه‌ی آن سند در یک سطر از فایل ورودی قرار داشت متفاوت است.

ابزار MALLET خروجی‌های متفاوتی را به کاربر ارائه می‌دهد. یکی از این خروجی‌ها میزان ارتباط بین برچسب و موضوع است که توسط عددی بین ۰ و ۱ مشخص می شود. هرچه مقدار این عدد بزرگ‌تر باشد بدان معناست که آن برچسب و موضوع به هم مرتبط‌ترند. در این حالت، خروجی مدل به‌صورت یک ماتریس ۱۰۰*۱۰۰ است به‌گونه‌ای‌که ستون‌های آن نشان‌دهنده‌ی موضوعات و سطرهای آن نشان‌دهنده‌ی برچسب‌ها است.

$$st_{ij} = P\left(topic_j \mid t_i\right) \quad (۷)$$

از خروجی‌های دیگری که ابزار MALLET در اختیار کاربر قرار می‌دهد این است که مرتبط‌ترین کلمات به هر موضوع را در قالب یک فایل خروجی مشخص می‌کند. تعداد مرتبط‌ترین کلمه به هر موضوع نیز توسط کاربر قابل‌تنظیم است.

**گام ۲: ترجمه‌ی کلمات پرس‌وجو به مجموعه‌ای از موضوعات**

پس‌ازآن‌که در گام اول میزان ارتباط بین هر برچسب و موضوع به‌دست آمد، برای هر برچسب مرتبط‌ترین موضوعات تا جایی انتخاب می‌شوند که مجموع مقدار ارتباط بین برچسب و موضوع بیشتر یا مساوی یک مقدارآستانه‌ی $\alpha$ شود.

$$\sum P\left(topic_j \mid t_i\right) >= \alpha \quad (۸)$$

**گام ۳: ترجمه‌ی کلمات پرس‌وجو به مجموعه‌ای از کلمات**

پس از ترجمه‌ی هر برچسب به مجموعه‌ای از موضوعات، با استفاده از رابطه (۹) سهم هر موضوع در مشخص‌شدن ترجمه‌ی یک برچسب محاسبه می‌شود.

$$ShareTopic = \left[ \frac{P(topic \mid t)}{\sum_{for\ all\ selected\ topic} P(topic \mid t)} * 10 \right] \quad (۹)$$

For each selected topic

رابطه‌ی (۹) سهم هر موضوع از ۱۰ ترجمه‌ای که برای هر برچسب باید انتخاب شوند را مشخص می‌کند. فرض کنید سهم یک موضوع در ترجمه‌ی یک برچسب، ۵ باشد در این حالت ۵ کلمه‌ی اول آن موضوع که درواقع بیشترین ارتباط با آن موضوع دارند به‌عنوان ترجمه‌ی آن برچسب درنظر گرفته می‌شود.

**گام چهارم: بازیابی و رتبه‌بندی افراد خبره**

پس از مشخص‌شدن ترجمه‌ی یک پرس‌وجو، باید با استفاده از آن‌ها افراد خبره شناسایی شوند. در بازیابی افراد خبره می‌توان سندها یا پاسخ‌هایی که کلمات ترجمه در آن‌ها به‌کاربرده‌شده را بازیابی کرد و به نویسنده‌ی آن جواب یک امتیاز داد. درنهایت نیز مجموع امتیازات را برای هر نویسنده محاسبه کرده و نویسنده‌ها براساس امتیازات کسب‌شده به صورت نزولی مرتب می‌شوند.

$$Score(e,t) = \sum_{\{d:e\in d, d\in R(Translation(t))\}} 1 \quad (۱۰)$$

در رابطه‌ی (۱۰) منظور از R(Translation(t)) مجموعه‌ی سندهایی است که حداقل یکی از کلمات ترجمه‌ی پرس‌وجو در آن‌ها به‌کار رفته باشد.

## ۳ -پیاده‌سازی روش پیشنهادی

در این بخش ابتدا به‌نحوه‌ی تولید مجموعه‌ی آزمون و پارامترهای آن پرداخته می‌شود. سپس پارامترهای دخیل در راه‌اندازی آزمایشات و معیار ارزیابی معرفی و نحوه‌ی مقداردهی آن‌ها بیان شده‌است. درنهایت نتایج حاصل از پیاده‌سازی مدل‌های پیشنهادی با روش‌های معروف موجود در این حوزه مقایسه شده‌است.

### ۳ ۱ -مجموعه‌ی آزمون

مدل‌های پیشنهادشده در این مقاله برروی مجموعه آزمون Stack Overflow که شامل ۲۴،۱۲۰،۵۲۳ پست در بازه‌ی زمانی آگوست ۲۰۰۸ تا مارچ ۲۰۱۵ است، اجرا و ارزیابی شده‌است. به‌منظور کاهش حجم مجموعه آزمون درنتیجه کاهش زمان اجرا، فقط سوالات با برچسب java به‌همراه مجموعه‌ی تمام جواب‌ها به‌عنوان مجموعه آزمون درنظر گرفته شده‌است. این مجموعه آزمون دارای ۰۷۱،۸۱۰ سوال و ۸۱۲،۵۱۰،۱جواب است که در مجموع شامل ۸۸۳،۳۲۰،۲ پست است. ازآنجاکه یک سوال در Stack Overflow می‌تواند چند برچسب داشته باشد، از مجموعه‌ی تمام برچسب‌هایی که در این پست‌ها همراه با جاوا آمده‌اند ۱۰۰ برچسب که بیشترین رخداد را با برچسب جاوا دارند به‌عنوان پرس‌وجو انتخاب شده‌اند. برای مشخص‌کردن خبره‌های واقعی در یک برچسب خاص، دو شرط باید به‌طور هم‌زمان اتفاق بیفتد. شرط اول این است که یک شخص باید بزرگ‌تر یا مساوری ۱۰ جواب پذیرفته‌شده داشته‌باشد [۲۶] و شرط دوم آن است که درصد جواب‌های پذیرفته‌شده‌ی یک خبره باید از میانگین درصد جواب‌های پذیرفته‌شده‌ی کل مجموعه‌ی آزمون بیش‌تر باشد [۲۷]. برای انجام عملیات مدل‌سازی موضوعی از ابزار MALLET استفاده شده‌است. MALLET یک بسته‌ی مبتنی‌بر جاوا است که در پردازش زبان طبیعی، دسته‌بندی و خوشه‌بندی سندها، مدل‌سازی موضوعی و استخراج اطلاعات کاربرد دارد.

### ۳ ۲ -معیار ارزیابی

برای ارزیابی مدل‌های پیشنهادی و مدل‌های مبنا از معیار ارزیابی MAP استفاده شده‌است. ازآنجاکه در سیستم‌های بازیابی افراد خبره تعداد خبره‌های هر پرس‌وجو در مجموعه‌ی آزمون مشخص است همچنین در این سیستم‌ها هر دو معیار دقت و فراخوانی مهم هستند درنتیجه معیار MAP به‌عنوان یک معیار ترکیبی که به‌طور هم‌زمان دقت و فراخوانی را در نظر می‌گیرد یکی از روش‌های مناسب جهت ارزیابی رتبه‌بندی‌های انجام‌شده است. روش محاسبه‌ی این معیار در رابطه‌ی زیر آورده شده‌است:

$$MAP = \frac{1}{|Q|}\sum_{q=1}^{|Q|}\left(\frac{1}{|\mathrm{Eq}|}\sum_{i=1}^{|R_q|} 1_{R_q^i \in E_q}\left(P@i\right)\right) \quad (۱۰)$$

که در رابطه‌ی (۱۱)، Q نشان‌دهنده‌ی مجموعه‌ی کل پرس‌وجوها در مجموعه‌ی آزمون، $E_q$ نشان‌دهنده‌ی مجموعه‌ی خبره‌های پرس‌وجوی q در مجموعه‌ی آزمون، $R_q$ نشان‌دهنده‌ی مجموعه‌ی خبره‌های شناسایی‌شده در پرس‌وجوی q با استفاده از روش‌های پیشنهادی، $1_{condition}=1$ اگر condition = true باشد و P@i نشان‌دهنده‌ی دقت سیستم بازیابی افراد خبره در خروجی شماره‌ی i سیستم است.

### ۳ ۳ -پارامترهای مسئله

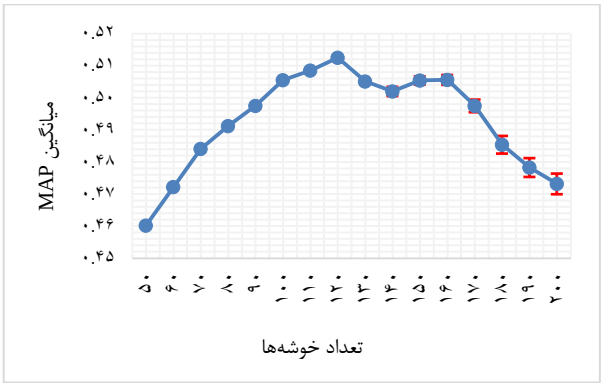
در این بخش پارامترهای دخیل در آزمایشات و نحوه‌ی مقداردهی آن‌ها بیان شده‌است. ازآنجاکه مدل یک مبتنی‌بر خوشه‌بندی است برای اندازه‌گیری دقیق MAP، تعداد خوشه‌ها در بازه‌ی ۵۰ تا ۲۰۰ با پرش ۱۰ تایی انتخاب و نتایج تحلیل شده‌است. ارزیابی مدل دو نیز با مقادیر مختلف پارامتر α به‌منظور تحلیل دقیق‌تر روش انجام شده‌است.

## ۴ -نتایج شبیه‌سازی و تحلیل نتایج

در مدل یک پیشنهادی از روش خوشه‌بندی K-Means، به‌دلیل پیاده‌سازی آسان و زمان اجرای کم آن، با معیار فاصله‌ی کسینوسی به‌منظور خوشه‌بندی نرمال‌شده‌ی بردارهای کلمات استفاده شده‌است. برای تخمین میانگین متوسط دقت، تعداد خوشه‌ها (K) در بازه‌ی ۵۰ تا ۲۰۰ با پرش‌های ۱۰تایی تغییر داده‌شده و به‌ازای هر تعداد خوشه، ۵۰ بار ترجمه‌ی پرس‌وجوها را به‌دست آورده و میانگین و واریانس MAP به‌ازای مقادیر مختلف K در جدول (۲) گزارش شده‌است. همان‌گونه‌که جدول (۲) نشان می‌دهد، بیشترین دقت مدل یک زمانی است که تعداد خوشه‌ها برابر با ۱۲۰ باشد. در شکل (۲)، کارایی مدل یک به‌ازای مقادیر مختلف K در قالب نمودار آورده شده‌است.

**جدول ۲: ارزیابی کارایی مدل یک با تعداد خوشه‌های متفاوت**

| تعداد خوشه | میانگین MAP | واریانس MAP |
|---|---|---|
| ۵۰ | ۰,۴۶۰۲ | ۰,۰۰۰۱۰۲۷۱۵ |
| ۶۰ | ۰,۴۷۲۲ | ۰,۰۰۰۱۳۳۷۵۴ |
| ۷۰ | ۰,۴۸۴۱ | ۰,۰۰۰۰۹۷۷۰۰ |
| ۸۰ | ۰,۴۹۱۲ | ۰,۰۰۰۰۶۵۲۵۶ |
| ۹۰ | ۰,۴۹۷۵ | ۰,۰۰۰۰۸۴۴۲۰ |
| ۱۰۰ | ۰,۵۰۵۵ | ۰,۰۰۰۰۷۹۷۰۰ |
| ۱۱۰ | ۰,۵۰۸۵ | ۰,۰۰۰۰۵۸۹۵۱ |
| **۱۲۰** | **۰,۵۱۲۵** | **۰,۰۰۰۰۵۵۳۶۷** |
| ۱۳۰ | ۰,۵۰۵۱ | ۰,۰۰۰۸۳۶۱۲۰ |
| ۱۴۰ | ۰,۵۰۲۰ | ۰,۰۰۱۴۳۱۶۸۷ |
| ۱۵۰ | ۰,۵۰۵۴ | ۰,۰۰۱۲۹۵۰۹۷ |
| ۱۶۰ | ۰,۵۰۵۶ | ۰,۰۰۱۵۳۷۶۸۵ |
| ۱۷۰ | ۰,۴۹۷۵ | ۰,۰۰۱۹۷۱۳۷۸ |
| ۱۸۰ | ۰,۴۸۵۴ | ۰,۰۰۲۷۴۹۰۶۸ |
| ۱۹۰ | ۰,۴۷۸۳ | ۰,۰۰۲۹۳۳۵۰۳ |
| ۲۰۰ | ۰,۴۷۳۲ | ۰,۰۰۳۲۲۰۴۳۶ |




**شکل ۲: نمودار کارایی مدل یک با تعداد خوشه‌های متفاوت**

در جدول (۳) نمونه‌ای از ترجمه‌های انجام‌شده توسط مدل یک آمده‌است. در این جدول سعی شده‌است تا با استفاده از کدگذاری رنگی برخی مفاهیم را به‌صورت راحت‌تری به خواننده منتقل نمود. رنگ زرد نشان‌دهنده آن است که کلمه‌ی زمینه‌ی خبرگی موردنظر دقیقا پیدا شده‌است. به‌عبارتی‌دیگر کلمه‌ی زمینه‌ی خبرگی موردنظر به خودش هم ترجمه شده‌است. رنگ سبز یعنی کلمه‌ی زمینه‌ی خبرگی به‌صورت یک زیررشته پیدا شده‌است. رنگ آبی بیان‌گر کلماتی است که ارتباط جالب و نزدیکی با زمینه‌ی خبرگی موردنظر دارند یا به‌عبارتی کلمات کلیدی برای آن زمینه‌ی خبرگی هستند. کلماتی که رنگ خاکستری دارند درواقع همان کلمه‌ی زمینه‌ی خبرگی هستند ولی در آن‌ها غلط املایی دیده می‌شود. همان‌گونه‌که نتایج حاصل از ترجمه نشان می‌دهد، کلمات پیشنهادی توسط روش یک پیشنهادی تا حد بالایی به کلمه پرس‌وجو مرتبط هستند و تا حد بسیارخوبی می‌توانند فاصله واژگانی بین کلمه پرس‌وجو و ترجمه‌ها را مدل نمایند. یک نکته قابل‌توجه در جدول (۳) این است که روش پیشنهادی در پاسخ به هر کلمه پرس‌وجو، نه‌تنها کلمات با املای مشابه را برگردانده‌است بلکه کلمات مرتبط دیگری که هر یک از جنبه‌های متفاوتی به کلمه پرس‌وجو نزدیک هستند و به‌گونه‌ای بیان‌گر خبرگی در زمینه کلمه پرس‌وجوی کاربر هستند نتیجه داده‌است. این‌نتیجه ناشی از خوشه‌بندی بردار کلمات است که با لحاظ‌کردن فاصله بردار کلمه به یک خوشه و فاصله بردار خوشه به کلمه پرس‌وجو سعی در رفع فاصله واژگانی دارد.

در جدول (۴) نمونه‌ای از ترجمه‌های انجام‌شده توسط مدل دو گزارش شده‌است. شکل (۴)، کارایی مدل پیشنهادی دو را به‌ازای مقادیر مختلف α در قالب نمودار نشان می‌دهد. بیشترین کارایی این مدل زمانی حاصل می‌شود که مقدار پارامتر α برابر با ۰٫۷ تنظیم گردد. نتایج جدول (۴) نیز برای مقدار پارامتر α برابر با ۰٫۷ گزارش شده‌است. همان‌گونه‌که مشاهده می‌شود این مدل نیز جواب‌های قابل‌قبولی در پاسخ به کلمه‌ی پرس‌وجوی کاربر برمی‌گرداند.

در جدول (۵) کارایی مدل‌های پیشنهادی در این مقاله با روش‌های ارائه‌شده در [۱، ۴، ۱۷] مقایسه شده‌است. همان‌طورکه مشاهده می‌شود مدل‌های پیشنهادی در این مقاله از نظر معیار ارزیابی MAP مدل‌های ارائه‌شده در [۱، ۴] را بهبود داده‌اند. مدل دوم پیشنهادی در این مقاله به‌لحاظ معیار ارزیابی MAP از روش اطلاعات متقابل که در [۱۷] ارائه‌شده بهتر است و مدل ۱ مقاله‌ی ما نیز از روش درج کلمه که در [۱۷] ارائه‌شده بهتر است.

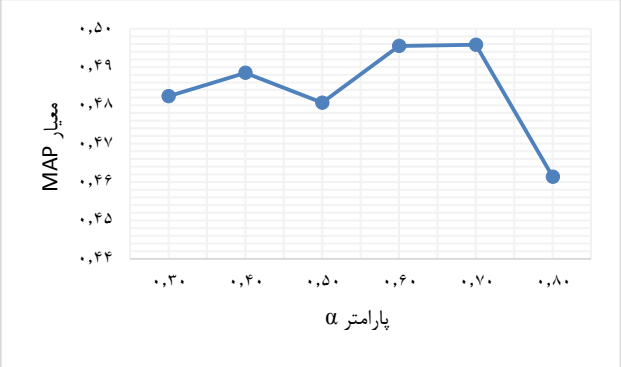
به‌منظور حصول اطمینان بیشتر از نتایج به‌دست‌آمده، در این مقاله از تست آماری تی وابسته استفاده شده‌است. در این تست تفاوت معنادار میانگین روش‌های پیشنهادی و روش‌های موردمقایسه به‌ازای مقادیر پیش‌فرض برای روش پیشنهادی با سطح اطمینان ۰٫۰۵ تست شده‌است. در جدول (۶) نتایج این آزمون بیان شده‌است. در این جدول عبارت x~y نشان‌دهنده آن است که میانگین دو روش x و y از لحاظ آماری تفاوت چندانی با هم ندارد و عبارت x<y بیان‌گر این است که میانگین روش y به طرز معناداری از میانگین روش x بیش‌تر است.

**جدول ۳: ترجمه‌ی چند کلمه پرس‌وجو نمونه توسط مدل یک**

| پرس‌وجو | ترجمه‌ی ۱ | ترجمه‌ی ۲ | ترجمه‌ی ۳ | ترجمه‌ی ۴ | ترجمه‌ی ۵ | ترجمه‌ی ۶ | ترجمه‌ی ۷ | ترجمه‌ی ۸ |
|---|---|---|---|---|---|---|---|---|
| **android** | developer.android.com | activity | android's | super.oncreate | tv.settext | savedinstancestate | r.layout.main | getapplicationcontext |
| **arrays** | arrays | array | arrays.copyof | myarray | dimensional | array.length | arrays.equals | numbers.length |
| **image** | srgb | saveimage | javax.imageio | showimage | imagereader | raster | originalheight | imageio |
| **hashmap** | map.size | map.entryset | map.keyset | map's | entry.getkey | hashmaps | entry.getvalue | e.getkey |
| **string** | substring | string's | s.length | str.trim | stringbuffer | str | str.substring | strings |
| **security** | secure | trust | certificates | malicious | tampering | insecure | certificate | compromised |
| **performance** | optimise | micro | optimizing | optimisation | warm | optimize | optimising | optimised |
| **mysql** | mysql | jdbc mysql | com.mysql.jdbc.driver | useunicode | mysql's | characterencoding | java.sql.preparedstatement | table_schema |
| **exception** | runtimeexception | throwing | catching | thrown | caught | catched | exceptions | throwable |
| **arraylist** | list.get | arraylist | list.add | arraylists | list.size | mylist.add | list.indexof | mylist.size |
| **sorting** | sort | ascending | sorting | descending | double.compare | integer.compare | java.util.comparator | compareto |
| **sql** | resultset | sql | stmt.close | sqlexception | connection.getmetadata | stmt.executequery | connection.createstatement | preparestatement |
| **python** | python | python's | jython | numpy | cpython | nltk | scipy | pythonic |
| **hibernate** | hibernate | hibernate's | tx.commit | org.hibernate.session | hibernateexception | dialect | hiberante | hibernate.connection.password |
| **hadoop** | hadoop | hdfs | hbase | mapred | hadoop's | site.xml | jobconf | namenode |
| **file** | test.txt | file.isfile | f.exists | filewriter | bufferedwriter | file.txt | file.listfiles | bw.close |
| **database** | resultset | sqlexception | stmt.close | sql | connection.createstatement | preparestatement | stmt.executequery | connection.preparestatement |
| **date** | week | simpledateformat | month | date | sdf | cal | gettime | simpledateformatter |
| **c#** | vb.net | net | system.linq | dotnet | system.collections.generic | console.writeline | xna | system.text |
| **algorithm** | iterative | Combinations | sqrt | subsets | breadth | visited | algorithm | shortest |
| **eclipse** | eclipse's | eclipse | indigo | workspace | helios | Kepler | assist | eclispe |
| **encryption** | encryption | encrypt | decrypt | encrypted | bouncy | keybytes | aeskey | decryption |
| **jdbc** | resultset | stmt.close | sqlexception | rs.close | stmt.executequery | connection.createstatement | executequery | preparestatement |
| **php** | php | file_get_contents | print_r | var_dump | php.ini | fclose | curlopt_returntransfer | php's |
| **jvm** | jvm | heap | jvms | hotspot | oome | generational | young | xmx |

**جدول ۴: ترجمه چند کلمه پرس‌وجو نمونه توسط مدل دو**

| پرس‌وجو | ترجمه‌ها |
|---|---|
| exception | exception, catch, exceptions, throw |
| awt | awt, graphics, swing, code, public, add, frame, import, swing |
| json | http, json, string, jsonobject, gson, public, class |
| image | code, public, image, int, bufferedimage, file, images |
| annotations | annotation, annotations, class, public, http, code, href, spring |
| struts2 | jsp, pre, struts, action, result, interceptor |
| encryption | key, byte, cipher, string, password, encryption, aes, encrypt |
| File-io | file, java, io, read, line, files, string |
| Spring-mvc | spring, servlet, mvc, web, code, spring |
| applet | applet, java, html, applets, java, file |




**شکل ۴: نمودار کارایی مدل دو به ازای مقادیر مختلف α**

در جدول (۷)، ترجمه‌های انجام‌شده توسط دو روش مطرح ارائه‌شده در[۱۷] و مدل‌های پیشنهادی یک و دو مقایسه شده‌است. همان‌طورکه مشخص است ترجمه‌های مدل‌های پیشنهادی در این مقاله نه تنها به کلمات پرس‌وجو بسیار مرتبط‌تر می‌باشند بلکه گستره‌ی بیش‌تری از خبرگی در حوزه کلمه پرس‌وجو را شامل می‌شوند. برای مثال در ترجمه‌ی کلمه‌ی پرس‌وجوی Array، روش‌های ارائه‌شده در [۱۷]، فاصله واژگانی را با پیشنهاد کلماتی از قبیل Array، Length و Int مدل کرده‌اند. این در حالی است که این کلمات چندان هم بیانگر خبرگی در حوزه‌ی آرایه نمی‌باشند. ازطرفی به نظر می رسد که مدل‌های پیشنهادی در این مقاله با پیشنهاد کلماتی از قبیل Array، Array.length، Arrays.copyof، Dimensional و Arrays.equals فاصله واژگانی بین کلمات پرس‌وجو و ترجمه‌ها را بهتر مدل کرده‌اند.

## ۵ نتیجه

فعالیت گسترده‌ی کاربران در فضای مجازی باعث افزایش روزافزون اطلاعات شده‌است که رسیدن به دانش را به امری دشوار مبدل کرده‌است. دراین‌میان سیستم‌های پاسخ به پرسش ازجمله سیستم‌هایی هستند که در جهت تسهیل دست‌یابی کاربر به دانش موردنظر مطرح شده‌اند. سیستم‌های بازیابی افراد خبره نیز از جمله سیستم‌هایی هستند که باتوجه به نیاز اطلاعاتی کاربران، متخصصین را که در واقع همان منابع دانش هستند به کاربر پیشنهاد می‌دهند. در این مقاله دو مدل ترجمه برای بازیابی و رتبه‌بندی افراد خبره ارائه شده‌است که مبتنی‌بر خوشه‌بندی و مدل‌سازی موضوعی بوده و در قیاس با مدل‌های زبانی و سایر مدل‌های ترجمه‌ی مطرح موجود نتایج بهتری ارائه داده‌است. مدل‌های پیشنهادی در این مقاله بر روی مجموعه‌ی پست‌های Stack Overflow به عنوان مجموعه‌ی آزمون تست و ارزیابی شده‌اند. نتایج به‌دست‌آمده نشان می‌دهد که مدل‌های ارائه‌شده در مقایسه با مدل‌های مرجع، دقت عملیات بازیابی و رتبه‌بندی افراد خبره را بهبود بخشیده‌اند.

**جدول ۵: مقایسه‌ی کارایی مدل‌های پیشنهادی با روش‌های موجود**

| روش | MAP |
|---|---|
| **مدل زبانی ۱ بلوگ [۳]** | ۰,۳۷۷۰ |
| **مدل زبانی ۲ بلوگ [۳]** | ۰,۳۶۲۰ |
| **مدل‌سازی موضوعی [۴]** | ۰,۴۳۴۰ |
| **اطلاعات متقابل [۱۷]** | ۰,۴۷۸۰ |
| **درج کلمه [۱۷]** | ۰,۴۹۶۰ |
| **مدل یک پیشنهادی**: مدل احتمالاتی مبتنی برخوشه بندی | ۰,۵۱۲۵ |
| **مدل دو پیشنهادی**: مدل مبتنی بر مدل‌سازی موضوعی | ۰,۴۹۵۸ |

**جدول ۶: مقایسه‌ی آماری روش‌های پیشنهادی و مدل‌های مبنا. $x \sim y$بیان‌گر عدم تفاوت معنادار میانگین دقت روش‌های $x$ و $y$ است.**

| روش‌های مورد مقایسه | آزمون آماری تی وابسته |
|---|---|
| مقایسه آماری مدل یک پیشنهادی و سایر مدل‌های مبنا | LM1 ~ LM2 < TM < MI < WE < P1 |
| مقایسه آماری مدل دو پیشنهادی و سایر مدل‌های مبنا | LM1 ~ LM2 < TM < MI < P2 < WE |
| مقایسه آماری دو مدل پیشنهادی | P2 < P1 |
| مدل زبانی ۱ بلوگ [۳] ≡ LM1، مدل زبانی ۲ بلوگ [۳] ≡ LM2، مدل‌سازی موضوعی [۴] ≡ TM | |
| اطلاعات متقابل [۱۷] ≡ MI، درج کلمه [۱۷] ≡ WE، مدل یک پیشنهادی ≡ P1، مدل دو پیشنهادی ≡ P2 | |

**جدول ۷: نمونه‌ای از ترجمه‌های انجام‌شده توسط روش‌های ارائه‌شده در[۱۷] و مدل‌های ارائه‌شده در این مقاله.**

| پرس‌وجو | نام روش | ترجمه۱ | ترجمه۲ | ترجمه۳ | ترجمه۴ | ترجمه۵ | ترجمه۶ |
|---|---|---|---|---|---|---|---|
| hibernate | اطلاعات متقابل[۱۷] | hibernate | entity | table | coulmn | sessionfactory | id |
| | درج کلمه[۱۷] | hibernate | entity | employee | table | query | jpa |
| | مدل یک | hibernate | hibernate's | tx.commit | org.hibernate.session | hibernateexception | dialect |
| | مدل دو | http | hibernate | session | id | org | jpa |
| swing | اطلاعات متقابل[۱۷] | textsample | jframe | jpanel | jbutton | swing | frame |
| | درج کلمه[۱۷] | jpanel | jbutton | jlabel | jframe | label | frame |
| | مدل یک | java.awt.event.actionlistener | actionperformed | javax.swing | javax.swing.jbutton | jbutton | actionlistener |
| | مدل دو | swing | text | add | frame | import | swing |
| selenium | اطلاعات متقابل[۱۷] | method.apply | selenium | webdriver | Driver.findelement | webelement | driver |
| | درج کلمه[۱۷] | tests | junit | test | mock | assertequals | unit |
| | مدل یک | selenium | webdriver | sendkeys | firefoxdriver | org.openqa.selenium.by | org.openqa.selenium.webdriver |
| | مدل دو | driver | selenium | element | webdriver | findelement | xpath |
| arrays | اطلاعات متقابل[۱۷] | array | int | 0 | arrays | 1 | j |
| | درج کلمه[۱۷] | array | index | int | System.out.println | arr | length |
| | مدل یک | arrays | array | arrays.copyof | myarray | dimensional | array.length |
| | مدل دو | array | int | length | arrays | string | code |

## مراجع


[۱] مریم باسره، ولی درهمی و سجاد ظریف‌زاده، «ارائه‌ی روشی برای استخراج خودکار عبارات کلیدی از اخبار وب پارسی»، مجله‌ی مهندسی برق دانشگاه تبریز، دوره‌ی ۴۷ شماره‌ی ۳، صفحه ۸۶۶-۸۵۵، ۱۳۹۶.

[۲] رضا خدایی، محمدعلی بالافر و سیدناصر رضوی، «اثربخشی بسط پرس‌وجو مبتنی بر خوشه‌بندی اسناد شبه‌بازخورد با الگوریتم KNN»، مجله‌ی مهندسی برق دانشگاه تبریز، دوره‌ی ۴۶ شماره‌ی ۱، صفحه ۱۵۱-۱۴۳، ۱۳۹۵.

[3] K. Balog, L. Azzopardi, and M. de Rijke. "A language modeling framework for expert finding." *Information Processing & Management*, vol. 45, no. 1, pp. 1-19, 2009.

[4] M. Karimzadehgan and Ch. Zhai, "Estimation of statistical translation models based on mutual information for ad hoc information retrieval", In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 323-330. ACM, 2010.

[5] H. Li, and J. Xu. "Semantic matching in search." *Foundations and Trends® in Information Retrieval, vol.* 7, no. 5, pp. 343-469, 2014.

[6] S. Momtazi, and F. Naumann. "Topic modeling for expert finding using latent Dirichlet allocation." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 3, no. 5, pp. 346-353, 2013.

[7] C. Van Gysel, M. de Rijke, and M. Worring, "Unsupervised, efficient and semantic expertise retrieval.", In Proceedings of the 25th International Conference on World Wide Web, pp. 1069-1079. International World Wide Web Conferences Steering Committee, 2016.

[8] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, "Expertise retrieval." Foundations and Trends® in Information Retrieval, vol. 6, no. 2–3, pp. 127-256, 2012.

[9] C. Macdonald, and I. Ounis. ''Voting for candidates: adapting data fusion techniques for an expert search task.'' In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pp. 387-396. ACM, 2006.

[10] M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, L. Zhao, and S. Ma. "Expansion-based technologies in finding relevant and new information: Thu trec 2002: Novelty track experiments." *NIST SPECIAL PUBLICATION SP*, no. 251, pp. 586-590, 2003.

[11] Y. Cao, J. Liu, S. Bao, and H. Li. "Research on Expert Search at Enterprise Track of TREC 2005." In TREC, 2005.

[12] R. M. Cooke, S. ElSaadany, and X. Huang. "On the performance of social network and likelihood-based expert weighting schemes.", Reliability Engineering & System Safety, vol. 93, no. 5, pp. 745-756, 2008.

[13] C. D. Manning and H. Schütze, "Foundations of statistical natural language processing.", Vol. 999, Cambridge: MIT press, 1999.

[14] T. Mueller-Prothmann and I. Finke, "SELaKT-Social Network Analysis as a Method for Expert Localisation and Sustainable Knowledge Transfer." J. UCS, vol. 10, no. 6, pp. 691-701, 2004.

[15] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation.", Journal of machine Learning research, vol. 3, no. Jan, pp. 993-1022, 2003.

[16] Gerard. Salton, A. Wong and C. Yang, "A vector space model for automatic indexing.", Communications of the ACM, vol. 18, no. 11, pp. 613-620, 1975.

[17] C. Van Gysel, M. de Rijke, and M. Worring. "Unsupervised, efficient and semantic expertise retrieval." In *Proceedings of the 25th International Conference on World Wide Web*, pp. 1069-1079, 2016.

[18] S. Patil, and K. Lee. "Detecting experts on Quora: by their activity, quality of answers, linguistic characteristics and temporal behaviors." *Social network analysis and mining* 6, no. 1, 2016.

[19] A. Dargahi Nobari, S. Sotudeh Gharebagh and M. Neshati, "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, 2017.

[20] A. Berger and J. Lafferty, "Information retrieval as statistical translation.", In ACM SIGIR Forum, vol. 51, no. 2, pp. 219-226. ACM, 2017.

[21] R. Jin, A. G. Hauptmann and C. Zhai, "Language model for information retrieval.". In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 42-48. ACM, 2002.

[22] V. Lavrenko, M. Choquette and W. B. Croft, "Cross-lingual relevance models.", In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 175-182. ACM, 2002.

[23] J. Nie, M. Simard, P. Isabelle and R. Durand, "Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the Web.", In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 74-81. ACM, 1999.

[24] J. Xu, R.Weischedel and C. Nguyen, "Evaluating a probabilistic model for cross-lingual information retrieval.", In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 105-110. ACM, 2001.

[25] V. Murdock and W. B. Croft, "Simple translation models for sentence retrieval in factoid question answering.", In Proceedings of the SIGIR-2004 Workshop on Information Retrieval For Question Answering (IR4QA), pp. 31-35. 2004.

[26] D. van Dijk, M. Tsagkias and M. de Rijke, "Early detection of topical expertise in community question answering.", In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 995-998. ACM, 2015.

[27] V. Dimitrova, T. Kuflik, D. Chin, F. Ricci, P. Dolog, and G. Houben, eds, "User Modeling, Adaptation and Personalization", 22nd International Conference, UMAP 2014, Aalborg, Denmark, July 7-11, 2014. Proceedings. Vol. 8538. Springer, 2014.

## زیرنویس‌ها

[1] Vocabulary Gap
[2] Translation Model
[3] Topic Modeling
[4] Test Collection
[5] Mean Average Precision
[6] Recall-based
[7] Generative Probabilistic Models
[8] Candidate Generation Model
[9] Voting Models
[10] Topic Modeling
[11] Vector Space Model
[12] Latent Dirichlet Allocation (LDA)
[13] Mutual Information (MI)
[14] Word Embedding
[15] Cross-Language Information Retrieval
[16] Sentence Retrieval
[17] Clustering
[18] Test Collection