# روش جدید متن‌کاوی برای استخراج اطلاعات زمینه کاربر به منظور بهبود رتبه‌بندی نتایج موتور جستجو

**جواد داودی مقدم*و علی احمدی**

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران

## چکیده

**یکی از بزرگ‌ترین مشکلات پیش‌روی موتورهای جستجو، رفع ابهاماتی است که در جستار کاربران وجود دارد. این ابهامات می‌تواند دلایل متعددی داشته باشد که از جمله آنها تعدد معانی و مفاهیم مرتبط با یک جستار یا کاربردهای مختلف آن جستار است. اگر موتور جستجو نتواند این ابهام را به شکل صحیح برطرف کند، در ارائه نتایج خود به کاربر دچار اختلال و خطا خواهد شد و نیاز کاربر را برطرف نخواهد کرد. این موضوع نقش مهمی در تعیین میزان کارایی موتور جستجو خواهد داشت. در این مقاله هدف آن است تا با جمع‌آوری اطلاعات زمینه کاربر در طول زمان، به تفسیر جستار کاربر کمک کرده و درنتیجهٔ آن رتبه‌بندی نتایج موتور جستجو را بهبود بخشیم. زمینه کاربر به هر اطلاعاتی گفته می‌شود که به شناخت ویژگی‌ها و خصوصیات کاربر کمک کند. در این مقاله متن صفحات وبی که کاربر از آن‌ها بازدید می‌کند، مورد پردازش قرار می‌گیرند تا مفاهیم اصلی و کلیدی آن‌ها استخراج شود. استخراج این مفاهیم (زمینه کاربر) که در سمت کاربر و بر روی سیستم وی اتفاق خواهد افتاد، با افزونه‌ای خواهد بود که به همین منظور تولید و بر روی مرورگر نصب می‌شود؛ سپس زمینه کاربر، در ساختاری خاص در سمت کاربر و برای هر کاربر به‌صورت خصوصی نگهداری می‌شوند. هنگامی که جستجویی انجام می‌شود (با توجه به خلاصه‌ای که موتور جستجو در ازای معرفی هر پیوند ارائه می‌دهد)، میزان شباهت نتایج موتور جستجو با زمینه کاربر مورد محاسبه قرار گرفته و به‌ازای هر نتیجه میزان شباهت آن با زمینه کاربر محاسبه می‌شود؛ سپس آن نتایجی به کاربر پیشنهاد می‌شوند (در مرورگر پررنگ می‌شوند) که با زمینهٔ وی تطبیق بیشتری داشته باشند. همان‌طورکه از نتایج آزمایش‌های پایان مقاله مشهود است، استفاده از زمینه کاربر در رتبه‌بندی نتایج موتور جستجو تاثیر قابل توجهی دارد. بررسی‌ها نشان می‌دهد که در ارائه ۱۰ نتیجه اول مربوط به ۳۰ جستار دارای ابهام، به طور میانگین روش پیشنهادی ۴۳٪ و موتور جستجوی گوگل ۱۶٪ از نتایج خود را مرتبط با مفهوم اصلی جستار مورد نظر ارائه کرده‌اند.**

**واژگان کلیدی: متن‌کاوی، بازیابی اطلاعات، زمینه کاربر، رتبه بندی نتایج موتور جستجو**

# A Novel Text Mining Method for User Context Extraction to Improve Search Engine Results Ranking

**Javad Davoudi Moghaddam* & Ali Ahmadi**

Computer Engineering Faculty, K. N. Toosi University of Technology, Tehran, Iran

## Abstract

**Today, the importance of text processing and its usages is well known among researchers and students. The amount of textual, documental materials increase day by day. So we need useful ways to save them and retrieve information from these materials. For example, search engines such as Google, Yahoo, Bing and etc. need to read so many web documents and retrieve the most similar ones to the user query. In this example, necessity of real time ability should be mentioned. Keyphrase extraction and some other fields like Information extraction, natural language processing, text summarization, query understanding, machine translation, and text similarity are subsets of text processing. So many efforts in text processing have been established, but there are still many open problems, especially in semantically document understanding**

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

**subjects. Although these subjects seem not to be very hard for humankind but they are very complex and confusing for a computer, because there is no standard structure to save documents so that computers be able to extract semantics and contents.**

**Document understanding and keyphrase extraction are some of the most important text processing goals. Many statistical and linguistic approaches are proposed in order to address these complex goals. Some methods work based on multi documents and some others on single document which all are generally more difficult than multi documents methods. Some methods use learning algorithms with training data and some others do not. Using natural language processing tools or resources -like ontologies- are effective ways to improve results, but these tools are not reliable for all languages. There are some articles for keyphrase extraction based on co-occurrence and also some statistical methods. Moreover, sometimes it is an important feature for a method to make real time outputs. Based on these characteristics, many approaches have been proposed in the literature.**

**In this paper, we present a new approach for keyphrase extraction from a single document. We present a language-independent approach based on combination of statistical information extracted from document and some logical rules named fundamental text rules. In this approach, there is no need to any natural language processing, nor to ontology and nor to any document corpus. We illustrate a real time method to understand each document focuses by extracting its phrases from segmented document without using any learning algorithm. Then, the Score for each phrase is calculated based on its occurrence and its related phrases occurrences. Then, fundamental text rules omit some phrases based on their scores and their places in text. Remained phrases shows the document focuses. Evaluation shows that our approach takes a high recall and precision in key phrase extraction with very good accuracy in text focuses understanding. These keyphrases extracted of a text presents the most important concepts of that text and it is used to retrieve documents in search engines more efficiently.**

**Keywords: text mining, information retrieval, user context, search engine results ranking**

## ۱- مقدمه

فهم منظور کاربر[1] یا اطلاعات مورد نیاز وی که با یک جستار[2] مشخص می‌شود، مدت‌هاست که به‌عنوان یک بخش حیاتی از بازیابی مؤثر اطلاعات شناخته می‌شود. عدم وجود ساختار برای ذخیره‌سازی منابع در اینترنت، بازیابی اطلاعات را با چالش‌ها و مشکلات بسیار عدیده و متنوعی روبرو می‌کند. برخلاف پیشرفت‌ها و پژوهش‌های فراوانی که در این خصوص صورت گرفته است، هنوز بسیاری از مسائل حل‌نشده باقی مانده‌اند [1]. هدف جستجوگرها به‌طور اصولی بالابردن دو عامل دقت[3] و فراخوانی[4] [2] است که به‌ترتیب نشان‌دهنده تعداد اسناد مرتبط بازیابی‌شده به تعداد کل اسناد بازیابی‌شده و تعداد اسناد مرتبط بازیابی‌شده به تعداد کل اسناد مرتبط بازیابی‌شده است. همچنین از عامل دیگری به نام معیار اف[5] برای سنجش میزان موفقیت در بازیابی استفاده می‌شود [3]. برای این منظور درک عمیق‌تر مفهوم جستار کاربر با استفاده از اطلاعات زمینه‌ای او برای بالابردن دقت جستجو و بازیابی در موتورهای جستجو مورد توجه قرار می‌گیرد.

جستار به عبارتی اطلاق می‌شود که برای جستجو در اختیار موتور جستجو قرار می‌گیرد. هدف جستار، ارائه یک نمایش جزئی و غیر مبهم از نیازهای کاربر است. فهم جستار دارای ابعاد گوناگونی است که از جمله آن‌ها می‌توان به بازیابی اطلاعات[6]، پردازش زبان طبیعی[7] و جستجوی وب اشاره کرد. بر اساس آمار منتشرشده، به‌طور میانگین تعداد کلمات هر جستار دو تا سه کلمه است. کم‌بودن تعداد کلمات در یک جستار باعث ایجاد ابهام در جستار و درنتیجه منجر به نتایج نامرتبط در بازیابی می‌شود [4]؛ همچنین اعلام شده است که بیش از نیمی از کاربران تنها به صفحه نخست نتایج بازیابی‌شده اکتفا می‌کنند و به صفحات بعد مراجعه نمی‌کنند. حتی گفته شده که هر کاربر به‌طور متوسط تنها به دو نتیجه از میان نتایج بازیابی‌شده سر می‌زند و از بررسی سایر نتایج منصرف می‌شود [5]. این مورد نیز به ضرورت رفع ابهام از جستار و ارائه نتایج دقیق در همان نتایج اولیه اشاره دارد. ضمن اینکه پیچیدگی‌های زبانی نیز به این ابهام افزوده می‌شوند. به‌عنوان مثال زبان فارسی، زبانی ادبی است

[1] User intent
[2] Search
[3] Precision
[4] Recall
[5] F_measure

[6] Information retrevial
[7] Natural language processing

که در آن کلمات دارای بار معنایی گوناگونی هستند که بیش‌تر کاربران نیز به آن توجهی ندارند. اگر ما اندکی در فهم جستار کاربر دچار خطا شویم، آنگاه نتایج بازیابی‌شده ممکن است، بسیار از هدف کاربر دور باشند. علاوه‌بر موارد بالا، برخی مشکلات دیگر در این قسمت عبارتند از: عدم وجود وردنت‌های[1] تخصصی در زمینه‌های مختلف برای فهم جستار، محتوای پویای وب که به‌راحتی قابل تغییر و ویرایش است و تغییر مداوم علائق افراد [6]. هرچند که در روش پیشنهادی مشکل عدم وجود وردنت‌ها و تغییر محتوای وب همچنان باقی است؛ اما الگوریتم می‌تواند تغییر علاقه افراد را تعقیب و خود را با زمینه جدید کاربر وفق دهد. این مورد به این صورت اتفاق می‌افتد که کاربر به‌مرور زمان و با تغییر علاقه خود، صفحات وبی را که مشاهده می‌کند تغییر می‌دهد. از آنجا که در روش پیشنهادی منبع استخراج زمینه کاربر، محتوای وب مورد رجوع وی خواهد بود، به مرور زمان مفاهیمی که از قبل به‌عنوان زمینه وی، ذخیره شده‌اند، در صفحات وب کمتر دیده می‌شوند و در فرآیند ارزش‌دهی در طول زمان تنزل پیدا می‌کنند. درعوض بر اساس صفحات وبی که بر اساس علاقه جدید مورد رجوع قرار می‌گیرند، مفاهیم جدیدی به زمینه وی افزوده می‌شوند که دارای ارزش بالاتری خواهند بود. بدین گونه زمینه کاربر در این مدل به‌طور کامل پویا و تطبیق‌پذیر با زمان طراحی شده است.

در ادامه ابتدا به تعریف و توصیف زمینه خواهیم پرداخت. در بخش سوم برجسته‌ترین کارهای مرتبط با پردازش جستار با استفاده از اطلاعات زمینه‌ای ارائه خواهد شد. بخش چهارم، منبع استخراج زمینه کاربر و روش پیشنهادی این پژوهش را در استفاده از این زمینه تشریح می‌کند و در انتهای این بخش نحوه استفاده از زمینه کاربر برای بهبود رتبه‌بندی نتایج موتور جستجو توضیح داده می‌شود. در بخش پنجم، نحوه آزمایش و ارزیابی روش پیشنهادی در استخراج زمینه کاربر و همچنین تأثیر آن بر بهبود رتبه‌بندی نتایج موتور جستجو ارائه شده است. بخش ششم نیز حاوی پیشنهادهایی برای کارهای آتی است.

## ۲- زمینهٔ کاربر

کلمهٔ زمینه[2] در اصطلاح به معنی قطعاتی از یک نوشته یا سخن است که پیش یا پس از یک کلمه خاص می‌آید و به برداشت معنی و مفهوم کلمه خاص کمک می‌کند. مطابق تعریف دی[3] «هر اطلاعاتی که به کار برده می‌شود تا موقعیت یک موجودیت را مشخص کند، زمینه است». بنابراین تعریف، زمینه به جنبه‌های خاصی محدود نمی‌شود. بر همین اساس زمینه به دو دسته عمده تقسیم می‌شود: زمینهٔ اولیه شامل اطلاعات محیطی مشترک یا اطلاعات ضروری، زمینهٔ ثانویه شامل اطلاعات شخصی کاربر. به‌عنوان مثال اگر هویت فرد را داشته باشیم، می‌توانیم شماره تلفن او را که زمینه ثانویه محسوب می‌شود، به‌دست آوریم [7]. محاسبات آگاه به زمینه برای نخستین‌بار توسط آقایان شیت و تیمر[4] در سال ۱۹۹۴ مطرح شد. آن‌ها محاسبات آگاه به زمینه را نرم‌افزارهایی تعریف کردند که خود را با زمینه‌های کاربر منطبق می‌ساختند. تعاریف زیر برای محاسبات آگاه به زمینه رایج است [8]:

- قابلیت دستگاه‌های محاسباتی برای کشف و احساس و تفسیر و پاسخ به جنبه‌های محیطی کاربر
- عملکرد خودکار سامانه نرم‌افزاری بر اساس محتوای کاربر
- تغییر و تطبیق پویا مبتنی بر محتوای برنامه و کاربر

در این مقاله، زمینهٔ کاربر به هر اطلاعاتی گفته می‌شود که به شناخت ویژگی‌ها، خصوصیات، عادت‌ها، رفتارها، حوزهٔ کاری و تخصصی و... کاربر کمک کند. از جمله منابعی که این اطلاعات را از آن می‌توان استخراج کرد و در این مقاله مورد توجه قرار گرفته است، متن صفحات وبی است که کاربر از آن‌ها بازدید می‌کند. این متون مورد پردازش قرار می‌گیرند تا مفاهیم اصلی و کلیدی آن‌ها استخراج و در ساختاری مشخص برای همان کاربر ذخیره شود.

## ۳- کارهای مرتبط

اغلب کارهای ارائه‌شده برای پردازش جستار، اطلاعات زمینه‌ای کاربران را برای تصمیم‌گیری‌های کلی مورد استفاده قرار می‌دهند. بدین صورت که با تجمیع و پردازش تمامی اطلاعات زمینه‌ای که از کاربران جمع‌آوری کرده‌اند، جهت‌گیری کلی را در جستجوها مشخص می‌کنند. ویژگی‌ها و اطلاعات به‌دست آمده از این پردازش در تعیین سیاست‌های کلی موتور جستجو دخالت داده می‌شود. آنچه که در این پژوهش به آن توجه شده است، استفاده از

[1] Wordnet
[2] Context
[3] Dey
[4] Theimer

اطلاعات زمینه‌ای کاربر برای بهبود نتایج موتور جستجو برای همان کاربر است. یعنی اگر دو کاربر با اطلاعات زمینه‌ای متفاوت یک عبارت را جستجو کنند، ترتیب نمایش نتایج برای هر کدام متناسب با اطلاعات زمینه‌ای خود فرد باشد. این امر باعث می‌شود که نتایج نمایش داده شده برای هر کاربر خصوصی‌سازی و دقیق‌تر شود. آنچه در این بخش به آن پرداخته می‌شود، کارهای مرتبطی است که از اطلاعات زمینه‌ای تجمیع‌شده استفاده می‌کنند. در بخش‌های بعد نحوهٔ استفاده از اطلاعات زمینه‌ای برای خصوصی‌سازی نتایج جستجو برای هر فرد بر اساس زمینهٔ وی تشریح خواهد شد.

## ۳-۱- طبقه‌بندی هدف جستار

در یک نوع تقسیم‌بندی که بر اساس اطلاعات زمینه‌ای کاربران به‌دست آمده است، اهداف جستجو در سه محور پیوندی[1]، اطلاعاتی[2] و منبع[3] خلاصه می‌شوند. این اهداف به‌صورت زیر تشریح می‌شوند:

- پیوندی: هدف کاربر یافتن پیوند یا نشانی خاصی بوده است.
- اطلاعاتی: هدف کاربر یافتن اطلاعات در مورد موضوعی خاص است.
- منبع: هدف کاربر دریافت یک منبع مانند فایل یا نرم‌افزار یا ... است.

اطلاعات زمینه‌ای استفاده‌شده در این پژوهش، نحوه عملکرد و پایگاه‌هایی بوده است که کاربران پس از جستجو به آن‌ها مراجعه کرده‌اند. بر اساس پژوهش‌های برنارد جانسن که بر روی بیست هزار نمونه جستار انجام شده است، ۵۱/۳٪ از جستارها در گروه اطلاعاتی، ۳۳/۵٪ جستارها در گروه پیوندی و ۱۵/۳٪ جستارها در گروه منبعی دسته‌بندی شده‌اند [6]. از جمله چالش‌های پیش روی این روش آن است که تقسیم‌بندی بالا بر اساس تجربه و بررسی جستارها صورت گرفته است و به‌صورت دقیق نمی‌توان ادعا کرد که این تقسیم‌بندی تمام جستارهای کاربران را تحت پوشش قرار می‌دهد. همچنین مرزبندی میان گروه‌ها و دسته‌های موجود در این تقسیم‌بندی واضح و روشن نیست و نمی‌توان ادعا کرد که یک جستار، تنها به یک دسته تعلق دارد.

## ۳-۲- تحلیل ساختاری و دستوری جستار

روش‌های متفاوتی برای این گونه پردازش جستار موجود است که در اینجا به برخی از آن‌ها اشاره می‌کنیم. برخی از روش‌ها تنها مبتنی بر تحلیل‌های زبانی و آماری هستند؛ [10]، [9] اما برخی دیگر در پردازش‌های خود از اطلاعات زمینه‌ای کاربران نیز استفاده می‌کنند.

برخی از پژوهش‌گران از روش جایگزینی برای تولید جستارهای جدید بهره می‌برند؛ بدین صورت که با تغییر برخی عبارات موجود در جستار، جستار جدیدی را تولید می‌کنند که ارتباط معنایی با جستار اصلی دارد. در این روش (جایگزینی یک عبارت با عبارت دیگر) جستارهای جدید دارای طول یکسان از نظر تعداد کلمات با جستارهای اصلی هستند. برای انتخاب عبارات جایگزین از اطلاعات جستجوهای قبلی استفاده می‌شود؛ که می‌توان آن‌ها را به‌عنوان اطلاعات زمینه‌ای جمع‌آوری‌شده از کاربران در نظر گرفت [11]. روشی که شرح داده شد جزو روش‌های پالایش تعاملی جستار است. یعنی برنامه با همکاری و نظر کاربر اقدام به پردازش جستار می‌کند و به‌صورت خودکار کاری را انجام نمی‌دهد. در روش دیگری که از پالایش تعاملی جستار استفاده می‌کند، پس از آنکه جستار کاربر تعیین شد، برخی مفاهیم که ارتباط آن‌ها با جستار کنونی بر اساس اطلاعات جستجوهای قبلی مشخص شده است به کاربر نمایش داده می‌شود. کاربر حق انتخاب دارد که از میان مفاهیم ارائه‌شده یک مفهوم را که ارتباط بیشتری با جستار کنونی دارد، انتخاب کند؛ سپس از این مفهوم و مطالب مرتبط با آن که از قبل شناخته‌شده هستند برای گسترش جستار اصلی استفاده می‌شود [12].

در پژوهشی دیگر که از همین روش استفاده می‌کند، در روشی جدید ابتدا میزان مشابهت جستار را با جستارهای گذشته محاسبه می‌کند. بر اساس این محاسبات جستارهایی که به جستار مورد نظر شباهت قابل قبولی داشته باشند، مبنای پردازش‌های بعدی قرار می‌گیرند؛ سپس این جستارهای مشابه را به‌عنوان پیشنهاد به کاربر معرفی می‌کند تا بدین وسیله بتواند با ترغیب کاربر به استفاده از این جستارهای پیشنهادی از میزان ابهام جستار قبلی بکاهد [13].

در موتورهای جستجو نشانی‌هایی که کاربران پس از انجام جستجو کلیک می‌کنند، می‌تواند به‌عنوان یک بازخورد ضمنی از موفقیت جستجو در نظر گرفته شود. این بازخوردها

[1] Navigational
[2] Informational
[3] Resource

کمک می‌کنند تا بتوان برای جستارهای مبهم بر اساس میزان شباهت با جستارهای قبلی نتایج مناسبی را بازیابی کرد یا برای پیشنهاد جستار به کاربر مورد استفاده قرارداد. برای مطالعه روش‌های بیشتر می‌توان به مراجع [14] و [15] مراجعه کرد.

## ۳-۳- شناخت ویژگی‌های پنهان جستار

برخی معتقدند که اگر بتوان با پردازش اطلاعات زمینه‌ای جمع‌آوری‌شده، ویژگی‌ها و ابعاد دیگری از جستار را شناسایی کرد، به گویاترشدن جستار کمک می‌شود و در نتیجه بازیابی اسناد به وسیله موتور جستجو دقیق‌تر خواهد شد. این ویژگی‌ها عبارت‌اند از: نوع جستار، عنوان جستار، هدف جستار، اختصاص جستار، دامنه جستار، حساسیت اعتبار، حساسیت‌های مکانی و زمانی [16].

بر اساس اطلاعاتی که از ثبت لاگ جستار به‌دست می‌آید، ویژگی‌های جستار را می‌توان استخراج کرد. به‌عنوان مثال طول جستار، تکرار جستار، تعداد صفحات مشاهده‌شده و طول جلسه جستجو را می‌توان استخراج کرد. همچنین از دیگر استفاده‌های لاگ جستار در فهم جستار می‌توان به طبقه‌بندی بر اساس هدف، معنا، مکان و زمان اشاره کرد. به‌طور معمول در لاگ جستار کاربران چهار دسته از اطلاعات موجود است که شامل اطلاعات مربوط به کاربر مانند نشانی آی‌پی[1]، اطلاعات مربوط به جستار مثل متن جستار و زمان جستجو، اطلاعات مربوط به نشانی‌های کلیک‌شده پس از جستجو و اطلاعات مربوط به نتایجی است که موتور جستجو به‌عنوان خروجی به کاربر نمایش می‌دهد. از فواید منحصربه‌فرد این روش آن است که جستار را از نظر حساسیت به مکان و زمان مورد تحلیل می‌توان قرار داد. برخی روش‌های آماری نیز مطرح شده‌اند که بر اساس تکرار جستار در روز، هفته و ماه می‌توانند حساسیت زمانی جستار را بر اساس اطلاعات لاگ جستار پیش‌بینی کنند [17].

البته یکی از مشکلاتی که همواره گریبان‌گیر دسته‌بندی جستار بوده است، هزینه پردازشی آن است. دسته‌بندی جستار باید به‌صورت آنی انجام گیرد تا کاربر متوجه این امر نشود؛ درصورتی‌که موتورهای جستجو روزانه تا چند صد میلیون جستار را پردازش می‌کنند. مشکل زمانی پیچیده‌تر می‌شود که در نظر داشته باشیم دسته‌بندی صفحات وب نیز در حال تغییر هستند. درنتیجه باید روشی برای دسته‌بندی جستار وجود داشته باشد که علاوه‌بر سرعت و دقت به تغییرات موجود در وب نیز توجه داشته باشد [18].

## ۴-۳- تولید گراف جریان جستار

گراف جریان جستار، نمایشی مجتمع از اطلاعات موجود در لاگ مربوط به جستار است. این گراف، اطلاعات موجود در جلسه‌های متفاوت جستجو را در یک گراف همگن مجتمع می‌کند. گره‌های این گراف جستارهای منحصربه‌فرد هستند و لبه از یک گره به گره دیگر وجود دارد؛ درصورتی‌که این دو جستار در یک جلسه جستجو به‌صورت متوالی مورد استفاده قرار گرفته باشند. از این گراف که پس از هر جستجو توسعه داده می‌شود، در تعیین جستارهای مرتبط با یک جستار خاص می‌توان استفاده کرد [19].

در پژوهشی دیگر برای تفسیر جستار از گرافی مشابه گراف معرفی‌شده در روش قبل استفاده می‌شود. در این روش هر جستار به همراه نتایجی که کاربر بر روی آنان کلیک می‌کند، ذخیره می‌شوند؛ سپس با استفاده از این داده یک گراف دوقسمتی ایجاد می‌شود. در یک قسمت به‌ازای هر جستار تمام نشانی‌هایی که کاربر بر روی آن‌ها کلیک کرده است، نگهداری می‌شوند (یک نشانی می‌تواند چندین پیوند از جستارهای مختلف داشته باشد) و در قسمت دیگر جستارهای مربوط به هر نشانی به آن متصل شده‌اند. از این گراف می‌توان فهمید که به‌ازای جستارهای مشابه چه صفحاتی مورد بازبینی قرار گرفته‌اند یا اینکه یک نشانی خاص بر اساس چه جستارهایی کلیک شده است. همچنین مفهوم جستار را با توجه به نشانی‌هایی که کلیک می‌شوند، می‌توان پیش‌بینی و از این پیش‌بینی در جستجوهای بعدی کاربر استفاده کرد [20]. برای آشنایی بیشتر با شیوه‌های استفاده از اطلاعات جریان جستار می‌توانید به منابع [21] و [22] مراجعه کنید.

# ۴- روش پیشنهادی

روش‌ها و شیوه‌های گوناگونی به‌منظور جمع‌آوری اطلاعات زمینه کاربران وجود دارد. همچنین بیان شد که زمینه کاربر شامل هرگونه اطلاعاتی می‌شود که در شناخت وی می‌تواند مؤثر باشد؛ لذا هر میزان که اطلاعات بیشتری راجع به کاربران جمع‌آوری شود، شناخت آن‌ها و درک درست از نیاز آن‌ها آسان‌تر خواهد شد. در اغلب مواقع روش‌هایی مانند

[1] IP address

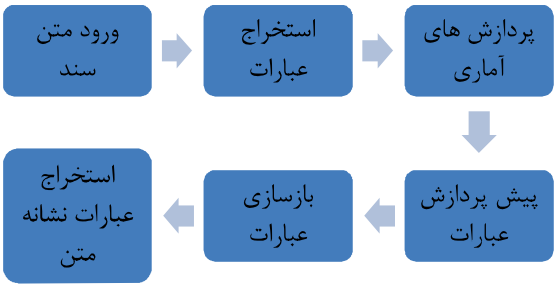
تکمیل فرم‌ها و اظهارنامه‌ها برای دریافت زمینه کاربر مورد استفاده قرار می‌گیرد. می‌توان ادعا کرد که این روش‌ها جزو روش‌های دقیقی هستند که در آن زمینه کاربر به‌صورت صریح از وی پرسیده می‌شود؛ اما امروزه مشکلاتی اساسی گریبان گیر این روش‌ها شده‌اند که از جمله می‌توان به‌این موارد اشاره کرد: تنوع و گستردگی و متغیربودن اطلاعات کاربران، خسته‌کننده و هزینه‌بربودن پرکردن اظهارنامه. با توجه به این مشکلات، لزوم استفاده از روش‌هایی که در آن زمینه کاربران به‌صورت خودکار و ماشینی استخراج شود تا هم نیاز به تکمیل اظهارنامه‌ها توسط کاربران نباشد و هم بتوان تغییرات زمینه‌ای کاربران را در طول زمان رصد کرد، احساس می‌شود. این‌چنین روش‌هایی به شناختی پایدار از افراد می‌توانند کمک کنند.

با توجه به توضیحات بالا، اساسی‌ترین نیاز را یافتن بستری مناسب می‌توان دانست که از طریق آن بتوان زمینه کاربر را تا حد قابل قبولی به‌دست آورد و البته تغییرات زمینه‌ای را در آن بستر رصد کرد. استخراج زمینه کاربر از سطح وب شامل موارد متعددی مانند اطلاعات موجود در صندوق‌های پستی، شبکه‌های اجتماعی و... است؛ به‌عنوان مثال در پژوهشی با استفاده از پیام‌های موجود در شبکه اجتماعی توییتر[1]، کلمات کلیدی استخراج می‌شود که به‌عنوان تگ[2] به کاربر تخصیص داده می‌شوند و کاربر با آن‌ها شناخته می‌شود [23]. در این روش‌ها سعی می‌شود با دسترسی به اطلاعات پروفایل شخصی کاربران در مورد آن‌ها اطلاعاتی کسب شود که بتوان برای تفسیر جستار از آن‌ها بهره جست. نمونه‌ای از این کارها را می‌توان در منابع [24] و [25] مشاهده کرد.

علاوه بر موارد بالا، یک مورد دیگر نیز وجود دارد که کمتر مورد توجه قرار گرفته است. این مورد همان صفحات وبی است که کاربر به‌صورت روزمره با آن‌ها سر و کار دارد. هر کدام از این سایت‌ها نشان‌دهنده برخی ویژگی‌های فرد می‌توانند باشند. برای استخراج زمینه کاربر از صفحات وب، نیاز است تا مفاهیم موجود در این صفحات استخراج شوند؛ سپس با گردآوری و یکپارچه‌سازی مفاهیم استخراج‌شده به زمینه فرد می‌توان پی برد؛ لذا استخراج دقیق و صحیح مفاهیم از صفحات وب یک نیاز اساسی در این پژوهش به حساب می‌آید. این پژوهش روش جدیدی را برای استخراج عبارات مهم از یک سند (برای مثال یک صفحه وب) ارائه می‌دهد. این روش مبتنی بر ترکیب اطلاعات آماری استخراج‌شده از سند و برخی قوانین است. این روش از هیچ گونه پردازش‌های زبان طبیعی استفاده نمی‌کند؛ لذا به هستان‌شناسی و یا دادگان آموزشی نیاز ندارد. ارزیابی‌هایی که در بخش بعد ارائه شده‌اند، نشان می‌دهد که استفاده از این روش نتایج بسیار خوبی را در استخراج عبارات مهم از متن داشته است. در ادامه این بخش ابتدا مراحل استخراج محتوا و مفاهیم از صفحات وب به‌طور دقیق شرح داده و سپس به نحوه استفاده از زمینه برای بهبود گویایی منظور جستار اشاره می‌شود.

## ۱-۴- فرآیند استخراج زمینه کاربر از صفحات وب

همان‌طورکه بیان شد یکی از منابع مهم برای استخراج زمینه کاربر صفحاتی است که کاربر در مرورگر خود، مورد بازدید قرار می‌دهد. هدف این است که در هر صفحه اینترنتی چند عبارت به‌عنوان **عبارات نشانه** که بیان‌گر مفاهیم اصلی آن صفحه بوده و نقاط تمرکز و تکیه متن را روشن می‌سازند، استخراج شوند. بدین‌وسیله با جمع‌آوری مجموع اطلاعاتی که هر شخص در طول زمان با آن‌ها سر و کار دارد، زمینه‌ای جامع را از وی می‌توان به‌دست آورد. هرچند اطلاعاتی که از این طریق کسب می‌شوند ممکن است چندان صریح، روشن وساخت‌یافته نباشند، اما درمجموع و با استفاده از روش‌های داده‌کاوی و پردازش متن کمک شایانی به شناخت زمینه کاربر می‌توانند بکنند. شکل (۱) مراحل به‌کاررفته را برای استخراج کلمات نشانه در روش پیشنهادی نشان می‌دهد.




**(شکل-۱): مراحل به‌کاررفته برای استخراج عبارات نشانه**
**(Figure-1): Keyphrases extraction steps**

### ۱-۱-۴- استخراج عبارات

در این مرحله متن خام صفحه وب به‌عنوان ورودی در نظر گرفته می‌شود. این متن خام متنی است که پس از تجزیه[3]

[1] Twiter
[2] Tag
[3] Parsing

صفحه وب اصلی و حذف مطالب اضافی شامل کدها و... به‌دست می‌آید. درواقع صفحه وب به‌صورت یک متن یکپارچه، ورودی این مرحله خواهد بود. این متن خام به‌صورت رشته‌ای از کلمات که به‌ترتیب در کنار یکدیگر قرار گرفته‌اند، در نظر گرفته می‌شود. این نکته بسیار اهمیت دارد که کلمات به‌صورت اتفاقی در این رشته به‌دنبال هم قرار نگرفته‌اند، بلکه حضور یک کلمه در مکانی خاص از رشته بیانگر معنای خاصی است که درصورت تغییر مکان آن، دیگر آن ارزش معنایی را نخواهد داشت. در بسیاری از روش‌های پردازش متن مانند روش نایو بیز[1] که به‌صورت آماری عمل می‌کنند، به مکان قرارگیری کلمات توجهی ندارند که این باعث می‌شود، بخش قابل‌توجهی از معانی و مفاهیم موجود در متن از بین برود. در این پژوهش با توجه به رویکرد پردازش معنایی متن، سعی بر این است تا کلمات و عبارات از بار معنایی جایگاه خود نیز بهره‌مند شوند. همان‌طورکه در انتهای قسمت ۴-۱-۳ اشاره خواهد شد، علاوه‌بر شاخص‌هایی که برای ارزش‌دهی به هر عبارت معرفی می‌شوند، سه قاعده نیز برای تصمیم‌گیری در مورد هر عبارت ارائه شده است. قواعد دوم و سوم از این مجموعه ناظر بر تعیین اهمیت یک عبارت بر اساس عبارات پیشین و پسین آن هستند. درواقع این قواعد برای توجه به ارزش مکان قرارگیری هر عبارت در متن، پیشنهاد شده‌اند.

برای استخراج عبارات متن، ابتدا رشته کلمات متن با علائم نگارشی از هم جدا می‌شوند و متن به مجموعه‌ای از زیررشته‌ها افراز می‌شود؛ سپس تمام عبارات تک‌کلمه‌ای تا سه‌کلمه‌ای که کلماتشان به‌ترتیب و پشت سر هم قرار گرفته‌اند، از هر زیررشته استخراج می‌شوند. دقت شود که قبل از این کار متن سند مورد نظر انتخاب بیشینه طول سه کلمه برای عبارات به‌دلایل زیر صورت گرفته است:

- بیش‌تر عبارات معنی‌دار در بازه یک تا سه کلمه قرار می‌گیرند. این ادعا بر اساس یک پردازش آماری است که روی ده‌هزار عبارت یک تا پنج کلمه‌ای انجام شده است. این ده‌هزار عبارت از تعدادی متن به روش بالا استخراج شده‌اند؛ سپس هر کدام از این عبارات با عامل انسانی در دو دستهٔ معنادار و بی‌معنی دسته‌بندی شده‌اند. نتیجه نشان داد که از بین ده‌هزار عبارت، ۶۳۰۰ عبارت معنادار هستند که بیش از ۹۲٪ از این عبارات معنادار دارای طول بین یک تا سه کلمه بودند.
- افزایش طول عبارات استخراج‌شده منجر به افزایش زمان پردازش می‌شود؛ یعنی هرچه میزان طول عبارات استخراج‌شده بیشتر باشد، زمان مورد نیاز برای استخراج عبارت نیز به‌صورت نمایی افزایش خواهد داشت؛ لذا بیشینه طول عبارات باید به‌گونه‌ای انتخاب شود که اطلاعات استخراج‌شده از عبارات طولانی‌تر، ارزش هزینه زمانی مذکور را داشته باشد. با توجه به اینکه از عبارات با طول بیشتر از سه، اطلاعات کمی به‌دست می‌آید، از استخراج آن‌ها صرف‌نظر شده است. به‌عنوان نمونه اگر "این متن، یک پژوهش علمی است؛ پژوهش علمی، نیاز به ارزیابی دارد." را به‌عنوان متن یک سند درنظر بگیریم، طبق جدول(۱) ابتدا این متن به چهار زیررشته تبدیل می‌شود و از هر کدام از زیررشته‌ها عبارت یک تا سه‌کلمه‌ای استخراج می‌شوند.

**(جدول-۱): استخراج عبارات یک تا سه کلمه از زیررشته‌های متن نمونه**

**(Table-1):Extraction of phrases up to three words from substrings of text**

| زیر رشته | عبارات استخراج شده |
|---|---|
| **این متن** | این-متن<br>این متن |
| **یک پژوهش علمی است** | یک-پژوهش-علمی-است<br>یک پژوهش-پژوهش علمی-علمی است<br>یک پژوهش علمی-پژوهش است |
| **پژوهش علمی** | پژوهش-علمی<br>پژوهش علمی |
| **نیاز به ارزیابی دارد** | نیاز-به-ارزیابی-دارد<br>نیاز به-به ارزیابی-ارزیابی دارد<br>نیاز به ارزیابی-به ارزیابی دارد |

### ۲-۱-۴- پیش‌پردازش عبارات

در این بخش به‌دنبال آن هستیم که عباراتی را که فاقد معنا و مفهوم معین و مهمی هستند، از دامنه عبارات موجود حذف کنیم. به‌همین‌منظور حذف ایست‌واژه‌ها[2] و عبارات بی‌معنی در این بخش صورت گرفت.

ایست‌واژه‌ها یا ایست‌کلمات به کلمات و عباراتی اطلاق می‌شود که به‌طورمعمول با تعداد تکرار بسیار زیاد در یک متن حضور دارند؛ اما از نظر مفهومی هیچ بار معنایی به‌خصوصی ندارند. از جمله این عبارات و کلمات می‌توان به

[1] Naïve bayes

[2] Stop word removal

ضمایر، قیدها، حروف ربط، حروف اضافه، نام‌آواها و... اشاره کرد. این عبارات بار مفهومی خاصی را در متن حمل نمی‌کنند و تنها از آن‌ها برای بیان بهتر جملات و ارتباط بین قسمت‌های متن استفاده می‌شود؛ لذا می‌توان ادعا کرد که حذف این عبارات از دامنه عبارات موجود، خدشه‌ای در درک مفهوم اصلی ایجاد نخواهد کرد. همچنین با بررسی‌های دقیق مشخص شد هر عبارتی که با ایست‌واژه‌ها شروع و یا به ایست‌واژه‌ها خاتمه یابد، به‌طورتقریبی در همه موارد فاقد یک معنای مشخص و معین است و از دامنه عبارات می‌تواند حذف شود. برای انجام این کار فهرستی از ایست‌واژه‌ها در این پژوهش فراهم شده است؛ لذا هر عبارت استخراج‌شده در مرحله قبل، اگر با ایست‌واژه شروع شوند یا خاتمه یابند، از دامنه عبارات حذف می‌شود. در این مقاله افعال نیز به‌عنوان ایست‌واژه در نظر گرفته شده‌اند. با اعمال این پیش‌پردازش جدول (۱) به جدول (۲) تبدیل خواهد شد.

**(جدول-۲): عبارات باقی‌مانده پس از مرحله پیش‌پردازش**

**(Table-2):Remaning phrases after preprocessing**

| زیر رشته | عبارات استخراج‌شده |
|---|---|
| **این متن** | متن<br>این متن |
| **یک پژوهش علمی است** | یک-پژوهش-علمی<br>یک پژوهش-پژوهش علمی<br>یک پژوهش علمی |
| **پژوهش علمی** | پژوهش-علمی<br>پژوهش علمی |
| **نیاز به ارزیابی دارد** | نیاز-ارزیابی<br>نیاز به ارزیابی |

### ۳-۱-۴- پردازش‌های آماری عبارات

این بخش به‌دنبال آن است تا با یک‌سری پردازش‌های آماری بر روی عبارات استخراج‌شده از متن که از پالایهٔ مرحله قبل عبور کرده‌اند، عباراتی را که به مفهوم متن نزدیک‌تر هستند، به‌عنوان عبارات نشانه انتخاب کند. علت استفاده از رویکرد آماری آن است که روش پیشنهادی، مستقل از زبان طراحی شده و برای هر متنی در هر زبانی قابل استفاده است. بسیار متداول است که یک فرد تنها به صفحات یک زبان رجوع نمی‌کند؛ بلکه به‌طورمعمول بیش از یک زبان در صفحات وب یک فرد دیده می‌شود. ازاین‌رو نیاز داریم تا پردازش مستقل از زبان بوده و یا ابزارهای زبانی را در زبان‌های متبوع در اختیار داشته باشیم.

نخستین شاخصه‌ای که باید برای هر عبارت محاسبه شود، تعداد تکرار آن عبارت در متن مورد نظر است. برای سهولت در ادامه این شاخصه را Document_Frequency نام‌گذاری می‌کنیم. برای یک عبارت مهم در متن که بار مفهومی خاصی را بر عهده دارد؛ بسیار دور از انتظار است که تنها یک‌بار در متن ظاهر شود؛ لذا اگر عبارتی به موضوعی خاص در متن اشاره داشته باشد انتظار می‌رود که بیش از یک‌بار در متن مورد استفاده قرار بگیرد. با این پیش‌فرض اگر عبارتی تنها یک‌بار در متن ظاهر شده باشد، حد آستانه شاخصه Document_Frequency را که عدد دو است، ارضا نمی‌کند و باید از دامنه عبارات فعال حذف شود.

شاخصهٔ دوم، تعداد زیر عبارت‌هایی برای هر عبارت است که آن زیرعبارت‌ها در بین عبارات فعال وجود داشته باشند. این شاخصه را Subphrase_Count می‌نامیم. این پارامتر نشان می‌دهد که تا چه میزان یک عبارت قابل اتکاست. در این پژوهش فرض بر این است که بین دو عبارت با طول مساوی، هر کدام که تعداد زیر عبارت‌های فعال بیشتری داشته باشد، احتمال حمل بار معنایی بیشتری دارد چرا که زیرعبارت‌های بیشتری از آن معنادار تشخیص داده شده‌اند. عبارات تک‌کلمه‌ای تنها دارای یک زیرعبارت هستند. اگر عبارتی دارای دو کلمه باشد. آنگاه سه زیرعبارت دارد که شامل یک عبارت دوکلمه‌ای و دو عبارت تک‌کلمه‌ای خواهد بود. این تعداد زیرعبارت‌ها با احتساب خود عبارت به‌عنوان بزرگ‌ترین زیرعبارت به‌دست می‌آید. توجه شود که زیرعبارت به‌ترتیبی از کلمات گفته می‌شود که در ترتیب کلمات عبارت اصلی ظاهر شود. حال اگر برخی از کلمات یک عبارت ایست‌واژه باشند، با توجه به مراحل قبل تعدادی از زیرعبارت‌ها از دامنه عبارات فعال حذف می‌شوند و تعداد آن‌ها کمتر از مقادیر معرفی‌شده خواهد شد. باید مقدار شاخصه Superstring_Count برای عبارت تک‌کلمه‌ای و دو کلمه‌ای به‌ترتیب دست‌کم دو و چهار باشد. اگر عبارتی این حد آستانه را برای شاخصه Superstring_Count ارضا نکند از دامنه عبارات فعال حذف می‌شود.

پارامتر سوم که Subphrase_Frequncy نام دارد، برابر مجموع تعداد تکرار زیر عبارت‌های هر عبارت است. این پارامتر معیار خوبی برای ارزش‌دهی به عبارت طولانی‌تر می‌تواند باشد؛ به‌خصوص عباراتی که در متن کمتر

ظاهرشده‌اند؛ ولی به‌سبب طول بیشتر دارای بار معنایی بیشتری هستند. اگر یک عبارت طولانی با تعداد تکرار کم، ارزش بالایی از نظر محتوایی در متن داشته باشد، انتظار می‌رود که زیرعبارت‌های آن، تعداد تکرار قابل قبولی را در متن داشته باشند. این پارامتر با درنظرگرفتن این فرضیه، به‌عبارت طولانی‌تر کمک می‌کند که در مقابل عبارت کوتاه‌تر اما با تکرار بیشتر، از میدان رقابت خارج نشود. درصورتی‌که اگر یک عبارت طولانی از نظر محتوایی بار چندانی را نداشته باشد، زیرعبارت‌های آن نیز در متن کمتر ظاهر می‌شوند. در این بخش برای عبارت تک‌کلمه‌ای حد آستانه در نظر گرفته نشده است. عبارت با دو کلمه دست‌کم باید دارای دو زیرعبارت در دامنه عبارات فعال باشد و عبارت با سه کلمه باید دست‌کم دارای چهار زیرعبارت در دامنه عبارات فعال باشد. اگر عبارتی یکی از شرایط بالا را نقض کند، یعنی حد آستانه شاخصه Document_Frequency را ارضا نکند، از دامنه عبارت فعال و سایر پردازش‌ها حذف می‌شود.

شاخصه چهارم با نام Substring_Score به محاسبه یک امتیاز برای هر زیررشته اختصاص دارد که این امتیاز به تک‌تک عبارت‌های موجود در آن زیررشته تخصیص می‌یابد. این امتیاز به‌صورت زیر محاسبه می‌شود:

(۱)

$$\text{Substring\_Scor} = \frac{\sum_{i=1}^{n} \text{Subphrase\_Frequncy}_i}{n}$$

در رابطه (۱)، $n$ تعداد عبارت‌های یک زیررشته و $\text{Subphrase\_Frequncy}_i$ بیان‌گر پارامتر Subphrase_Frequncy برای عبارت $i$ام در زیررشته است.

یک سؤال اساسی آن است که آیا یک عبارت مهم در یک محدوده کم اهمیت یا بی‌اهمیت در متن می‌تواند قرار بگیرد؟ یا به‌عبارت دیگر آیا یک زیررشته (جمله یا پاراگراف) با عبارات کم‌اهمیت یا بی‌اهمیت دربردارنده یک عبارت مهم می‌تواند باشد؟ به‌طورمعمول بین عبارات موجود در متن یک ارتباط و همبستگی وجود دارد. برخلاف تصویر که می‌تواند تغییرات ناگهانی و زیادی بین پیکسل‌های مجاور داشته باشد، عبارات متن اغلب چنین رفتاری را نشان نمی‌دهند. هرچند که متن فراز و نشیب‌های زیادی را از نظر مفهومی می‌تواند داشته باشد؛ اما باید دقت شود که این فراز و نشیب‌ها در بیش‌تر مواقع نوسان زیادی ندارند؛ به تعبیر دیگر عبارات مهم اغلب به‌صورت ناگهانی در متن ظاهر نمی‌توانند شوند؛ لذا چند قاعده طراحی شد که با استفاده از آن‌ها دامنه عبارات فعال باز هم محدودتر می‌شود:

I. اگر عبارتی وجود داشته باشد که شاخصه سوم و چهارم آن به‌ترتیب از میانگین کل شاخصه‌های سوم و چهارم عبارات متن کمتر باشد، آنگاه آن عبارت حذف خواهد شد.

II. اگر عبارتی وجود داشته باشند که به‌عنوان مثال تا چهار عبارت قبل یا بعد از آن در دامنه عبارات فعال حضورنداشته باشند، باعث می‌شود که مقادیر شاخصه‌های Subphrase_Frequncy و Substring_Score مربوط به عبارات حذف‌شده صفر در نظر گرفته شود؛ لذا محدوده اطراف آن عبارت کم‌اهمیت تشخیص داده می‌شود و درنتیجه خود آن عبارت نیز از دامنه عبارات فعال حذف خواهد شد.

III. اگر عبارتی وجود داشته باشد که مقادیر پارمترهای Subphrase_Frequncy و Substring_Score چندین عبارت قبل و بعد از آن کمتر از یک حد آستانه باشند، محدوده اطراف آن عبارت کم‌اهمیت تشخیص داده می‌شود و درنتیجه خود آن عبارت نیز از دامنه عبارات فعال حذف خواهد شد.

درنهایت و با اعمال شروط بالا بر روی دامنه عبارات فعال، عباراتی باقی خواهند ماند که نمایندگان خوبی برای بیان نکات مهم در متن می‌توانند باشند. توجه شود که آنچه در این پژوهش مورد نظر است، با واژگان کلیدی که در مقالات ارائه می‌شود، متفاوت است. واژگان کلیدی در مقالات ناظر بر موضوع کلی مقاله و جهت‌گیری اصلی مقاله است؛ اما در این پژوهش به‌دنبال آن هستیم تا عبارت‌هایی را پیدا کنیم که نشان می‌دهند در متن به چه موضوعاتی پرداخته شده است. می‌توان فرض کرد که واژگان کلیدی که در مقالات ارائه می‌شوند، زیرمجموعه‌ای از عبارات انتخاب‌شده می‌توانند باشند. در مرحله بعد عبارتی از دامنه حذف نمی‌شود، بلکه ممکن است دو یا چند عبارت باهم ترکیب شوند و عبارت بزرگ‌تری را تشکیل دهند که به‌تفصیل به آن می‌پردازیم. شاخصه‌های محاسبه‌شده برای عبارات جدول (۲) در جدول (۳) ارائه شده است. بر اساس اطلاعات (جدول میانگین پارامتر سوم ۲/۹ و میانگین پارامتر چهارم ۳/۸۷ می‌شود. بنابراین مواردی که به رنگ قرمز مشخص شده‌اند، بر اساس قاعده شماره یک حذف

می‌شوند. قـواعد ۲ و ۳ عـبارتی را برای این نـمونه حذف نمی‌کنند.

**(جدول-۳) : پارامترهای محاسبه‌شده برای عبارات متن نمونه**
**(Table-3): Parameter values calculated for sample text phrases**

| عبارت | Document frequency | Subphrase count | Subphrase frequency | Substring score |
|---|---|---|---|---|
| متن | 1 | 1 | 1 | 0.5 |
| یک | 1 | 1 | 1 | 5.75 |
| پژوهش | 2 | 1 | 2 | 5.75 |
| علمی | 2 | 1 | 1 | 5.75 |
| نیاز | 1 | 1 | 1 | 1.25 |
| ارزیابی | 1 | 1 | 1 | 1.25 |
| یک پژوهش | 1 | 3 | 4 | 5.75 |
| پژوهش علمی | 2 | 3 | 6 | 5.75 |
| یک پژوهش علمی | 1 | 6 | 9 | 5.75 |
| نیاز به ارزیابی | 1 | 3 | 3 | 1.25 |

### ۴-۱-۴- بازسازی عبارات

برای یافتن عبارات نشانه در متن، تمام عبارت‌هایی که تعداد کلمات کمتر یا مساوی سه دارند، استخراج و سپس با پردازش‌های بیان‌شده تعدادی به‌عنوان نماینده انتخاب می‌شوند. هرچند برای انتخاب عبارات با طول بیشینهٔ سه کلمه دلایلی بیان شد، اما پرواضح است که امکان وجود عبارت‌های طولانی‌تر که حامل پیام مهمی از متن باشند نیز دور از انتظار نیست؛ به‌گونه‌ای که وجود عبارت‌های کوتاه‌تر در فهرست عبارات نهایی شاید نتواند پوشش دقیق و کاملی از عبارات طولانی‌تر را به دست دهد. به همین منظور این بخش به‌عنوان بخش پایانی در استخراج عبارات نشانه افزوده شده است. در این بخش هر جفت از عبارات مهم انتخاب‌شده در مرحله قبل مورد بررسی قرار می‌گیرند. اگر ترکیب آن‌ها (قرار گرفتن آن‌ها پشت سر هم) باعث تولید عبارتی شود که در متن وجود دارد آن دو عبارت باهم ترکیب می‌شوند و تشکیل یک عبارت جدید با تعداد کلمات بیشتر را می‌دهند. این عمل ترکیب تا آنجایی پیش می‌رود که دیگر هیچ ترکیبی به وجود نیاید. عبارت‌هایی که در پایان این مرحله باقی می‌مانند، به‌عنوان زمینه کاربر در یک فایل ذخیره می‌شوند.

در جدول (۳) که نتیجه گام قبل را نشان می‌دهد، تنها سه عبارت "یک پژوهش"، "پژوهش علمی" و "یک پژوهش علمی" باقی مانده‌اند. دو عبارت نخست را بنا به موقعیت مکانی آنها در متن می‌شود ترکیب کرد که عبارت معتبر موجود در متن را می‌سازد. به‌صورت اتفاقی این عبارت ترکیبی همان عبارت سوم موجود در جدول (۳) است. بنابراین از کل متن نمونه مورد نظر تنها یک عبارت که همان "یک پژوهش علمی" است، باقی می‌ماند و تنها همین عبارت به‌عنوان عبارت نشانه متن نمونه انتخاب می‌شود.

## ۴-۲- استفاده از زمینه کاربر در بهبود رتبه‌بندی نتایج موتور جستجو

همان‌طورکه درقبل بیان شد، یکی از منابعی که دراختیارداشتن آن به شناخت ویژگی‌های یک فرد و کشف زمینه او می‌تواند کمک کند، مشخصات صفحات وبی است که وی به آن‌ها سر می‌زند. در طول زمان و با ثبت اطلاعات تعداد زیادی صفحات وب که کاربر مشاهده می‌کند، تا حد زیادی به ویژگی‌های او می‌توان پی برد. این روش علاوه‌بر ایجاد قابلیت یادگیری در طول زمان، تغییرات ویژگی‌های کاربر را نیز می‌تواند رصد کند. یک نکته بسیار مهم آن است که این عامل یادگیر یک عامل سمت کاربر[1] است. یعنی این عامل به‌گونه‌ای طراحی شده است که اطلاعات زمینه‌ای کاربر را بر روی سیستم خود کاربر ذخیره کند. این نوع رویکرد دو امتیاز مهم دارد که عبارت‌اند از: حفظ محرمانگی اطلاعات کاربر و جلوگیری از بار پردازشی زیاد در سمت سرور میزبان. البته برخی روش‌ها نیز وجود دارند که مایل به استفاده از رفتار کاربر در سمت میزبان هستند. با توجه به محدودیت‌هایی که برای استفاده از زمینه کاربر در سمت میزبان وجود دارد، برخی پژوهش‌ها اطلاعات کاربران را در حد جلسهٔ جستجو محدود کرده‌اند. جلسهٔ جستجو به بازهٔ زمانی پیوسته‌ای گفته می‌شود که کاربر در آن در حال جستجو است. در این شرایط برای جستار اولیه کاربر هیچ اطلاعاتی از رفتار وی در دسترس نیست؛ اما از پردازش جستارهای دوم به بعد می‌توان از رفتار کاربر پس از مشاهده نتایج جستارهای قبلی استفاده کرد. اینکه کاربر به‌ازای هر جستار بر روی چه نتایجی کلیک می‌کند، می‌تواند معنی و مفهوم خاصی داشته باشد که از آن بتوان برای جستجوهای بعدی استفاده کرد. دقت شود که در این حالت اطلاعات فرد پایدار نیست و با اتمام جلسهٔ جستجو اطلاعات مربوط به وی از بین می‌رود [26].

[1] Client side

حال به این سؤال پاسخ می‌دهیم که پردازش‌های روش پیشنهادی که در سمت کاربر انجام می‌شوند، چگونه بر روی نتایج موتور جستجو که در سمت میزبان قرار دارد، تأثیر می‌گذارند؟ روند طراحی‌شده بدین صورت است که موتور جستجو با توجه به جستاری که کاربر در اختیار وی قرار می‌دهد نتایجی را برای وی بازیابی می‌کند. برنامه پیشنهادی به‌صورت یک افزونه طراحی شده است که بر روی مرورگرهای مختلف قابل نصب باشد. زمانی‌که موتور جستجو نتایج را بر روی مرورگر کاربر نمایش می‌دهد به‌صورت آنی این نتایج به‌وسیلهٔ افزونه مورد پردازش قرار می‌گیرند؛ یعنی میزان مشابهت خلاصه نتایج نمایش داده‌شده در مرورگر با زمینه کاربر (که درقبل در سمت کاربر استخراج و ذخیره شده است) مورد انطباق و سنجش قرار می‌گیرد. درواقع افزونه، نتایج بازیابی موتور جستجو را به دو دسته نتایج مرتبط و نتایج غیرمرتبط با زمینه تقسیم می‌کند، سپس آن‌هایی را که با زمینه مرتبط تشخیص داده شده‌اند، به‌صورت مجزایی به‌عنوان مثال با رنگ خاص یا علامت ستاره نمایش می‌دهد. بدین ترتیب از زمینه کاربر برای خصوصی‌سازی نتایج بازیابی‌شده توسط موتور جستجو که در سمت میزبان فراهم شده است، استفاده می‌شود. با این رویکرد اگر دو کاربر با زمینه‌های گوناگون یک جستار را جستجو کنند که در حوزه تخصصی هر کدام معنا و مفهوم خاصی دارد، نتایج نشان‌داده‌شده به آن‌ها متفاوت خواهد بود و هر کدام منطبق با زمینهٔ خود نتایج مرتبط را خواهند دید.

## ۵- ارزیابی نتایج تجربی

هدف از این بخش بررسی و آزمون روش ارائه‌شده برای استخراج زمینه و ارزیابی نتایج آن در بهبود عملیات جستجوست. از آنجا که در این پژوهش روش جدیدی برای استخراج عبارات نشانه از متن پیشنهاد شده است، ابتدا نتایج آزمایش جهت ارزیابی دقت الگوریتم استخراج عبارات نشانه ارائه می‌شود و در ادامه هدف اصلی این پژوهش که همان استفاده از اطلاعات زمینه برای بهبود نتایج جستجو است، مورد ارزیابی قرار می‌گیرد.

### ۵-۱- ارزیابی طبقه‌بندی متون خبری

برای اطمینان از دقت و صحت روش پیشنهادی در استخراج عبارات نشانه، ارزیابی طبقه‌بندی متون خبری نیز مورد توجه قرار گرفت. یکی از منابع بسیار معتبر در طبقه‌بندی اسناد و متون، پایگاه داده خبری رویترز است. این پایگاه داده با جمع‌آوری بیست هزار خبر طبقه‌بندی‌شده جزو بزرگ‌ترین و مهم‌ترین پایگاه‌های داده مورد استفاده در پردازش متن است. این پایگاه داده شامل بیست گروه خبری مختلف است که هر کدام دارای هزار خبر مرتبط با آن گروه است [27]. از بین بیست گروه خبری موجود در این پایگاه داده، شش گروه خبری به تصادف انتخاب شدند و برای هر کدام از گروه‌های خبری نیز ۱۱۵ خبر به تصادف برگزیده شدند. این مجموعه که شامل شش گروه خبری و درجمع ۶۹۰ خبر می‌شد، برای ارزیابی این قسمت مد نظر قرار گرفت. شش گروه خبری انتخاب‌شده عبارت‌اند از:

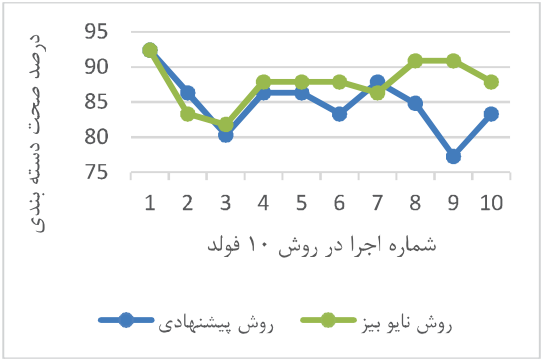
- Comp.graphics
- Misc.forsale
- Rec.autos
- Sci.crypt
- Soc.religion.christian
- Taik.politics.guns

نحوه ارزیابی و دسته‌بندی متون این گروه‌های خبری بر اساس روش سنجش متقاطع $k$-دسته‌ای[1] تعیین شد. در مرحله نخست داده‌ها به ده قسمت تقسیم شدند و برنامه نیز برای ده بار به اجرا گذاشته شد. در هر بار اجرا یک قسمت از ده قسمت به‌عنوان داده آزمایشی و نه قسمت دیگر به‌عنوان داده‌های آموزشی فرض می‌شدند. روال انجام ارزیابی نیز به این صورت بود که داده‌های آموزشی به‌عنوان ورودی به الگوریتم پیشنهادی داده می‌شدند و این الگوریتم، عباراتی را به‌عنوان نمایندگان این نه قسمت انتخاب می‌کرد. پس از پردازش داده‌های آموزشی تمامی گروه‌ها و تعیین نمایندگان هر گروه خبری، نوبت به دسته‌بندی داده‌های آزمایشی می‌رسید. هر کدام از داده‌های آزمایشی نیز با استفاده از الگوریتم پیشنهادی به تعدادی عبارت که نمایندگان هر داده آزمایشی بودند، تبدیل می‌شدند؛ سپس میزان شباهت عبارات هر داده آزمایشی با هر کدام از دسته‌ها بر اساس معیار مشابهت کسینوسی محاسبه می‌شد. داده آزمایشی متعلق به دسته‌ای قلمداد می‌شد که بیش‌ترین درصد شباهت را با آن داشته باشد.

در مرحله دوم نیز داده‌ها به پنج قسمت تقسیم شدند و دوباره فرآیند بالا به‌منظور دسته‌بندی متون خبری بر روی این پنج قسمت اجرا شد. درصد صحت دسته‌بندی متون خبری در هر کدام از اجراهای مرحله نخست و دوم به‌ترتیب در شکل‌های (۲ و ۳) نمایش داده شده است.

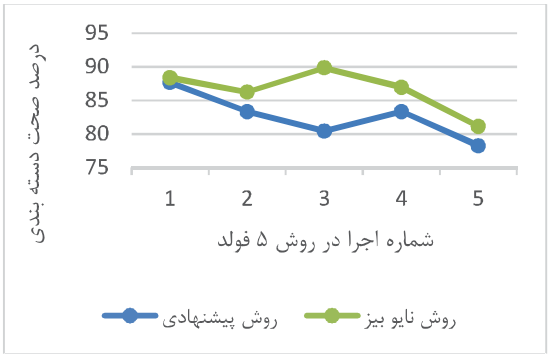
[1] K fold cross validation

میانگین صحت دسته‌بندی متون خبری که از مرحله نخست و مرحله دوم به‌دست آمده‌اند، در جدول (۴) نیز ارائه شده است. همان‌طورکه در جدول (۴) مشاهده می‌شود، درصدهای روش پیشنهادی بسیار نزدیک به روش نایو بیز است درصورتی‌که در روش نایو بیز تعداد کلمات استفاده شده به‌منظور دسته‌بندی، بسیار بیشتر از روش پیشنهادی بوده است. درنهایت نیز تفاوت میانگین درصد دسته‌بندی بین روش پیشنهادی و روش نایو بیز به‌طورتقریبی سه درصد است. این امر نشان می‌دهد که الگوریتم در انتخاب عبارت مهم و نشانه دقت بالایی دارد.




**(شکل-۲): مقایسه درصد صحت دسته‌بندی متون خبری در روش پیشنهادی با روش نایو بیز در ارزیابی ده فولد**

**(Figure-2): 10 fold naive bayes correctness of news texts classification vs proposed method compression**




**(شکل-۳): مقایسه درصد صحت دسته‌بندی متون خبری در روش پیشنهادی با روش نایو بیز در ارزیابی پنج فولد**

**(Figure-3): 5 fold naive bayes correctness of news texts classification vs proposed method compression**

**(جدول-۴): میانگین درصد صحت دسته‌بندی متون ۲۰ گروه خبری پایگاه داده رویترز**

**(Table-4): Average correctness of classification of 20 news groups of Reuter**

| | روش پیشنهادی | روش نایو بیز |
|---|---|---|
| میانگین ۱۰ فولد | 84.8% | 87.7% |
| میانگین ۵ فولد | 82.6% | 86.5% |
| میانگین کلی | 83.7% | 87.1% |

اگر میانگین کلی به‌دست‌آمده از این شش دسته خبری را نماینده‌ای از کل مجموعه بیست گروه خبری پایگاه داده رویترز در نظر بگیریم، می‌توانیم یک مقایسه نسبی بین روش پیشنهادی با روش‌های ارائه‌شده در سایر مقالات مطابق جدول (۵) ارائه دهیم. دقت شود که هدف از ارائه اطلاعات جدول (۵) مقایسه مستقیم روش پیشنهادی با سایر روش‌ها نیست (چون برای این منظور باید کلیه شرایط از جمله انتخاب داده‌های آموزشی و ارزشیابی به‌صورت یکسان انتخاب شود) بلکه هدف این است که نشان دهیم روش پیشنهادی در انتخاب عبارات نشانه از متون دقت خوبی را دارد. درصورتی‌که در روش نایو بیز تعداد کلمات بسیار بیشتری (به‌طورتقریبی سه برابر روش پیشنهادی) برای دسته‌بندی استفاده شده است.

**(جدول-۵): درصد صحت دسته‌بندی الگوریتم‌های مختلف بر روی متون ۲۰ گروه خبری پایگاه داده رویترز**

**(Table-5): Average correctness of classification of 20 news groups of Reuter in other methods**

| نام الگوریتم | درصد صحت دسته‌بندی |
|---|---|
| نایو بیز | 87.1% |
| DF+SVM [28] | 84.71% |
| PCA + SVM [28] | 83.97% |
| LDA + SVM [28] | 88.96% |
| روش پیشنهادی (۶ گروه خبری) | 83.7% |

## ۲-۵- ارزیابی بهبود رتبه‌بندی نتایج موتور جستجو با استفاده از زمینه کاربر

در این ارزیابی به‌دنبال آن بودیم تا بدانیم روش پیشنهادی چه تأثیری در بهبود عملیات جستجو دارد؛ لذا با توجه به روشی که در بخش ۴-۲ توضیح داده شد، برنامه پیشنهادی در کنار موتور جستجو قرار گرفت تا از بین نتایجی که موتور جستجو بازیابی می‌کند، آن‌هایی که به زمینه کاربر نزدیک‌تر هستند به کاربر پیشنهاد داده شوند. برنامه مذکور به‌صورت افزونه‌ای[1] بر روی مرورگر نصب می‌شود تا علاوه‌بر جمع‌آوری محتوای وب کاربر برای استخراج زمینه، در هنگام جستجو و ارائه نتایج به‌وسیلهٔ موتور جستجو اعلام نظر کند. برای انجام این ارزیابی در مرحله نخست نیاز بود تا بر اساس تاریخچه مراجعات کاربر به صفحات وب، زمینه او استخراج شود. برای

[1] Add-on

این کار از یک کاربر نمونه خواسته شد که صفحات خاصی را از ویکی‌پدیا که در رده‌بندی مشخصی قرار می‌گیرند، به‌کرّار با مرورگر مرور کند. ویکی‌پدیا بزرگ‌ترین مرجع دانشی چندزبانهٔ جهان است که به‌صورت رایگان بر روی وب قابل‌دسترسی است. نکته مهمی که ما را به استفاده از ویکی‌پدیا برای تولید زمینه ترغیب کرد، این بود که هر مقاله در ویکی‌پدیا شامل تعدادی رده است. رده در مقالات ویکی‌پدیا نشان می‌دهد که آن مقاله با چه مفاهیمی در ارتباط است. به‌عنوان مثال مقاله‌ای که در مورد درخت صحبت می‌کند در رده‌های مختلفی قرار می‌گیرد که یکی از آن‌ها رده گیاهان است؛ چون درخت در زمره گیاهان طبقه‌بندی می‌شود. به همین صورت برای تمامی مقالات موجود در ویکی‌پدیا یک یا چند رده وجود دارد که می‌توان حوزه مفهومی مقاله را با استفاده از رده‌های آن تخمین زد. برای این آزمایش مقالاتی از ویکی‌پدیا توسط کاربر مورد بازدید قرار گرفتند که دارای رده رایانه باشد. به این ترتیب با انتخاب این صفحات رایانه‌ای مجموعه‌ای صفحه وب داشتیم که در مورد مسائل مربوط به رایانه در آن‌ها بحث شده است. بنابراین عبارات نشانه به عنوان زمینه کاربر با استفاده از روش پیشنهادی از این مقالات استخراج شدند.

در حین استخراج زمینه باید به این سؤال پاسخ می‌دادیم که چه تعداد صفحه برای ساخت زمینه کاربر مناسب است. بر اساس پژوهش‌های به‌عمل‌آمده در سال ۲۰۰۴ نشان داده شده که دویست متن برای بیان جنبه‌های مختلف یک عبارت بهینه است [29]. در پژوهش‌های دیگر نیز که برای رفع ابهام از جستار وجود دارد، این عدد به‌عنوان یک مبنا در نظر گرفته شده است [30]. ما نیز از همین نتیجه استفاده می‌کنیم و زمینه کاربر را حداکثر با دویست صفحه نمایش می‌دهیم.

در مرحله بعد نیاز بود تا جستارهایی با مفاهیم چندگانه تولید شوند که موتور جستجو نتواند نتایج به‌طورکامل یک‌دستی را برای آن‌ها بازیابی کند. درواقع ما به دنبال یافتن جستارهایی بودیم که موتور جستجو در بازیابی اسناد آن‌ها دچار مشکل است و نمی‌داند که به‌طوردقیق کدام مفهوم از جستار مد نظر کاربر است. در سال ۲۰۰۵ رقابتی به‌منظور دسته‌بندی جستار با پشتیبانی ACM برگزار شد. داده‌هایی که سایت مسابقات در اختیار قرار داده بود؛ شامل هشتصدهزار جستار و ۶۷ گروه بودند. هدف این بود که برای هر جستار، از بین ۶۷ گروه معرفی‌شده گروه‌هایی که مرتبط با جستار است، تشخیص داده شوند. هر جستار غالباً در گروه‌های مختلف قرار می‌گرفت؛ لذا باید روشی پیشنهاد می‌شد که در آن تمام گروه‌هایی که یک جستار به آن‌ها مربوط می‌شود، استخراج شوند [31]. با توجه به اینکه جستارهای مورد استفاده در این مسابقه، دارای مفاهیم چندگانه بودند، ما از آن‌ها برای آزمایش خود استفاده کردیم. نمونه‌ای از جستارهای مبهم به‌همراه زمینه مربوطه در جدول (۶) نشان داده شده است. به‌عنوان‌مثال نخستین جستار مبهم در جدول (۶)، "روش‌های تشخیص نوفه" است. این جستار مبهم است چون روش‌های تشخیص نوفه در علوم و موضوعات گوناگونی کاربرد دارد؛ اما اگر بدانیم کاربری که آن را جستجو می‌کند در حوزه "پردازش تصویر" پژوهش و کار می‌کند، ابهام جستار برطرف می‌شود. درواقع در این پژوهش از زمینه کاربر برای رفع ابهام جستار استفاده می‌شود؛ اما روش‌های دیگری از جمله استفاده از اصلاح نامه‌ها در رفع ابهام عبارات مطرح هستند که از جمله کارهای صورت‌گرفته را در این رابطه به زبان فارسی را در مرجع [32] می‌توان مطالعه کرد.

به‌همین منظور سی جستار مبهم (دارای مفاهیم چندگانه) از منبع معرفی‌شده انتخاب شد. به‌ازای هر جستار مبهم دویست صفحه از ویکی‌پدیا استخراج شد که این دویست صفحه به‌عنوان زمینه و برای رفع ابهام از جستار مربوطه، مورد استفاده قرار گرفت. درواقع این دویست صفحه، اطلاعاتی را در برداشتند که ناظر بر یکی از مفاهیم جستار مبهم بود.

**(جدول -۶): جستارهای مبهم به همراه زمینه‌هایی برای رفع ابهام**
**(Table-6): Using context for ambiguous queries**

| Ambiguous query | Context |
|---|---|
| Noise detection methods | Image |
| configure wlan | Linux |
| Network configuration setting | Ubuntu |
| New version of operating system | Android |

در این سناریو، جستار کاربر بدون هیچ‌گونه تغییری در اختیار موتور جستجو قرار گرفت. زمانی که موتور جستجو نتایجش را ارائه داد، نوبت به پردازش‌های روش پیشنهادی می‌رسد تا از بین نتایج، آن‌هایی را که با زمینه کاربر مطابقت بیشتری داشته باشند مشخص کند. در این پژوهش موتور جستجوی گوگل[1] برای بازیابی نتایج و مقایسه مورد استفاده قرار گرفت. به این صورت که به‌ازای هر جستارده نتیجهٔ نخست گوگل به‌عنوان نتایج موتور جستجو در نظر گرفته شد؛ سپس از بین پنجاه نتیجه نخست موتور جستجو، ده

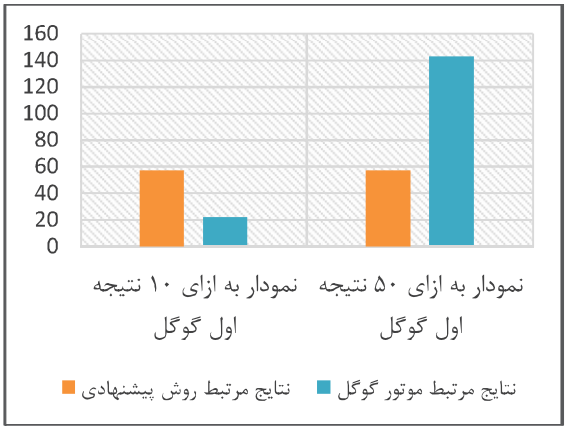
[1] Google.com

نتیجه بر اساس بیش‌ترین میزان شباهت با زمینه کاربر، انتخاب و به‌عنوان نتایج پیشنهادی ارائه شدند. مشابهت نتایج حاصله از موتور جستجو و نتایج ارائه‌شده توسط روش پیشنهادی با استفاده از معیار مشابهت‌یابی کسینوسی[1] به‌صورت زیر به‌دست می‌آید.

$$\text{similarity} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \qquad (۲)$$

در رابطه (۲)، $n$ کل عبارات متن نخست و دوم است. همچنین شاخص‌های $A_i$ و $B_i$ به ترتیب بیان‌کننده تعداد تکرار عبارت $i$ام در متن نخست و متن دوم هستند. در این قسمت متن نخست زمینه کاربر محسوب می‌شود که دربرگیرنده عبارات نشانه و متن دوم خلاصه‌ای است که موتور جستجو به‌ازای هر نتیجه بازیابی‌شده ارائه می‌دهد؛ سپس برای نتایج، موتور جستجو و روش پیشنهادی بررسی می‌شود که چه تعداد نتایج مرتبط با زمینه بوده‌اند. نتیجه این بررسی نشان می‌دهد که از مجموع سی ارزیابی به‌عمل‌آمده، در ۲۱ مورد روش پیشنهادی منجر به بهبود نتایج جستجو شده است. در شش مورد نتایج روش پیشنهادی با نتایج موتور جستجو یکسان است و سه مورد روش پیشنهادی از موتور جستجو ضعیف‌تر عمل کرده است.

نمودار تجمیعی نتایج که در شکل (۴) نمایش داده شده است، نشان می‌دهد که در ارائه ده نتیجه نخست مربوط به سی جستار دارای ابهام، به‌طور میانگین روش پیشنهادی ۴۳٪ و موتور جستجوی گوگل ۱۶٪ از نتایج خود را مرتبط با مفهوم اصلی جستار مورد نظر ارائه کرده‌اند. می‌توان نتیجه گرفت که روش پیشنهادی توانسته است با استفاده از زمینه، به بهبود گویایی منظور جستار کمک کند. همچنین مقایسه تعداد کل نتایج مرتبط روش پیشنهادی با کل نتایج مرتبط در پنجاه نتیجه نخست موتور جستجو نیز در شکل (۴) نشان داده شده است. مطابق این نتیجه هنوز هم نتایج مرتبطی وجود دارند که روش پیشنهادی نتوانسته است آن‌ها را تشخیص دهد. از جمله دلایل این اتفاق آن است که نتایج موتور جستجو در روش پیشنهادی تنها بر اساس خلاصه‌ای که ارائه می‌شوند، اولویت‌بندی می‌شوند و اگر دسترسی به متن صفحات نتایج جستجو ایجاد شود، تعداد بسیار بیشتری از نتایج مرتبط به‌وسیله روش پیشنهادی شناسایی خواهند شد. شبه‌کد روش پیشنهادی در این مقاله در شکل (۵) نیز ارائه شده است.




**(شکل-۴): نمودار تجمیعی تعداد نتایج مرتبط در موتور گوگل و روش پیشنهادی**

**(Figure-4):Google count of context related results vs proposed method**

۱. شروع
۲. اگر مرورگر صفحه وبی را بارگزاری کرد:
  ۲,۱. استخراج متن صفحه وب از فرمت HTML
  ۲,۲. استخراج تمامی عبارات از یک تا K کلمه از متن صفحه وب
  ۲,۳. حذف ایست کلمات و عبارات بی‌معنی از بین عبارات استخراج شده
  ۲,۴. تعیین مقدار هر کدام از پارامترهای زیر برای هر عبارت است(پردازش‌های آماری):
    ۲,۴,۱. Document_Frequency
    ۲,۴,۲. Subphrase_Count
    ۲,۴,۳. Subphrase_Frequncy
    ۲,۴,۴. Substring_Score
  ۲,۵. حذف عبارات کم اهمیت بر اساس قواعد سه گانه‌ای که بر روی ۴ پارامتر فوق اعمال می‌شوند
  ۲,۶. بازسازی عبارات نشانه
  ۲,۷. ثبت عبارات نشانه به عنوان زمینه کاربر در ساختاری مناسب
  ۲,۸. برو به شروع
۳. اگر کاربر عبارتی را در موتور جستجوی اینترنتی مورد جستجو قرار داد:
  ۳,۱. تعیین میزان شباهت کسینوسی عنوان و خلاصه ارائه شده به وسیله موتور جستجو برای هر لینک با زمینه کاربر
  ۳,۲. رتبه‌بندی مجدد نتایج بر اساس مقدار مشابهت کسینوسی
  ۳,۳. پیشنهاد نتایج مرتبط به زمینه کاربر از بین نتایج بازیابی شده به وسیله موتور جستجو
  ۳,۴. برو به شروع

**(شکل-۵): شبه‌کد روش پیشنهادی**

**(Figure-5): Pseudo code of proposed method**

[1] Cosine similarity

## ۶- نتیجه‌گیری و پیشنهادها

استخراج زمینه کاربر بر اساس صفحات وبی که مشاهده می‌کند، ایده اصلی این پژوهش بوده است، یعنی ویژگی‌های فردی، علاقه، سلیقه و... کاربر را از این داده‌ها می‌توان استخراج کرد؛ سپس از این اطلاعات برای بهبود رتبه‌بندی نتایج موتور جستجو بهره گرفته شده است. ارزیابی‌های به‌عمل‌آمده نشان می‌دهند که روش پیشنهادی دقت و صحت خوبی در انتخاب عبارات مهم از متن و همچنین بهبود رتبه‌بندی نتایج موتور جستجو دارد. استخراج زمینه، نگهداری داده‌ها و کلیه پردازش‌های مربوط به تقویت نتایج جستجوهای موتور جستجو همگی در سمت کاربر اتفاق می‌افتد. با توجه به شناختی که در این پژوهش در حوزه پردازش جستار و به‌خصوص پردازش جستار به‌وسیله زمینه کاربر به دست آمده است، پیشنهادهای زیر برای توسعه کارهای آینده مفید خواهد بود: خوشه‌بندی یا دسته‌بندی عبارت استخراج‌شده به‌عنوان زمینه، استفاده از هستان‌شناسی‌ها برای یافتن مفاهیم مرتبط با زمینه، استفاده از واژه‌نامه‌ها به‌منظور ترجمه زمینه کاربر زبان‌های دیگر، به‌روزرسانی زمینه و تغییر اطلاعات زمینه مطابق با تغییر رفتار کاربر.

## تشکر و قدردانی

از جناب آقای مهندس مصلی‌نژاد که با پیشنهادها، نظرات و همکاری خود در مراحل مختلف نگارش این مقاله به بهبود کیفیت آن کمک کردند، کمال تشکر و قدردانی را داریم.

## ۷- مراجع 7-References

[1] Hamdi, Mohamed Salah, "SOMSE: A semantic map based meta-search engine for the purpose of web information customization", Applied Soft Computing, vol. 11, no. 1, pp. 1310-1321, 2011.

[2] Mangold, Christoph, "A survey and classification of semantic search approaches", International Journal of Metadata, Semantics and Ontologies, vol. 2, no. 1, pp. 23-34, 2007.

[3] Kirar, Dilip, and Pranita Jain, "Equirs: Explicitly query understanding information retrieval system based on hmm", International Journal of Engineering Inventions, vol. 2, no 1, pp. 31-36, Jan. 2013.

[4] Vaughan, Liwen, and Mike Thelwall, "Search engine coverage bias: evidence and possible causes", Information processing & management, vol. 40, no. 4, pp. 693-707, 2004 .

[5] Jansen, Bernard J., et al, "Real life information retrieval: A study of user queries on the web", In ACM SIGIR Forum, vol. 32, no. 1, pp. 5-17, 1998 .

[6] Jansen, Bernard J., and Danielle Booth, "Classifying web queries by topic and user intent", CHI'10 Extended Abstracts on Human Factors in Computing Systems, pp. 4285-4290, ACM, 2010.

[7] Calderón-Benavides, Liliana, Cristina González-Caro, and Ricardo Baeza-Yates, "Towards a deeper understanding of the user's query intent", SIGIR 2010 Workshop on Query Representation and Understanding, pp. 21-24, 2010.

[8] Abowd, Gregory D., et al, "Towards a Better Understanding of Context and Context-Awareness", Handheld and ubiquitous computing. Springer Berlin Heidelberg, pp. 304-307, 1999 .

[9] Allan, James, and Hema Raghavan, "Using part-of-speech patterns to reduce query ambiguity", In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 307-314, 2002.

[10] Bäurle, Florian, "A user interface for semantic full text search", Master Thesis, Faculty of Engineerin, University of Freiburg, 2011.

[11] Bing, Lidong, and Wai Lam, "Investigation of web query refinement via Topic Analysis and Learning with Personalization", 2011.

[12] Fonseca, Bruno M., et al, "Concept-based interactive query expansion", In Proceedings of the 14th ACM international conference on Information and knowledge management, pp. 696-703, 2005 .

[13] Song, Wei, et al, "An effective query recommendation approach using semantic strateg-ies for intelligent information retrieval", Expert Systems with Applications, vol. 41, no. 2, pp. 366-372, 2014 .

[14] Bordogna, Gloria, et al, "Disambiguated query suggestions and personalized content-similarity and novelty ranking of clustered results to optimize web searches", Information Processing & Management, vol. 48, no. 3, pp. 419-437, 2012 .

[15] Broccolo, Daniele, et al, "Generating suggestions for queries in the long tail with an inverted index", Information Processing & Management, vol. 48, no. 2, pp. 326-339, 2012 .

[16] González-Caro, Cristina, and Ricardo Baeza-Yates, "A multi-faceted approach to query intent classification", String Processing and Informa-

tion Retrieval, Springer Berlin Heidelb-erg, pp. 368-379, 2011.

[17] Jiang, Daxin, Jian Pei, and Hang Li, "Mining Search and Browse Logs for Web Search: A Survey", ACM Transactions on Computational Logic, pp. 1-42, Apr. 2013.

[18] Li, Lin, et al, "A feature-free search query classification approach using semantic distance", Expert Systems with Applications, vol. 39, no. 12, pp. 10739-10748, 2012 .

[19] Bai, Lu, et al, "Exploring the query-flow graph with a mixture model for query recommendation", Proceedings of IGIR Work-shop on Query Representation and Understand-ing, Beijing, China, Jul. 2011.

[20] Beeferman, Doug, and Adam Berger, "Agglomerative clustering of a search engine query log." In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discov-ery and data mining, pp. 407-416, 2000 .

[21] Andersen, Casper, and Daniel Christensen, "User Logs for Query Disambiguation", 2013 .

[22] Sondhi, Parikshit, Raman Chandrasekar, and Robert Rounthwaite. "Using query context models to construct topical search engin-es", In Proceed-ings of the third symposium on Information interaction in context, pp. 75-84, 2010 .

[23] Wu, Wei, Bin Zhang, and Mari Ostendorf. "Automatic generation of personalized annotation tags for twitter users", In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 689-692, 2010 .

[24] Biancalana, Claudio, and Alessandro Micarelli. "Social tagging in query expansion: A new way for personalized web search", In Computational Science and Engineering, vol. 4, pp. 1060-1065, 2009 .

[25] Kramár, Tomáš, Michal Barla, and Mária Bieliková. "Disambiguating search by leveraging a social context based on the stream of user's activity", In User Modeling, Adaptation, and Personalization, pp. 387-392, 2010 .

[26] Cao, Huanhuan, et al, "Context-aware query classification", In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 3-10, 2009 .

[27] Joachims, Thorsten. "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization", Carnegie-mellon univ Pittsburgh pa dept of computer science, No. CMU-CS-96-118, 1996 .

[28] Luo, Le, and Li Li. "Defining and evaluating classification algorithm for high-dimensional data based on latent topics", PloS one 9, No. 1, 2014 .

[29] Zeng, Hua-Jun, et al, "Learning to cluster web search results", In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 210-217, 2004 .

[30] Song, Ruihua, et al, "Identification of ambiguous queries in web search", Information Processing & Management, vol. 45, no. 2, pp. 216-229, 2009 .

[31] Li, Ying, Zijian Zheng, and Honghua Kathy Dai, "KDD CUP-2005 report: Facing a great challenge", ACM SIGKDD Explorations Newsletter 7, no. 2, pp. 91-99, 2005.

[۳۲] فرهاد راد، حمید پروین، آتوسا دهباشی و بهروز مینایی. "ارائه روشی جدید برای شاخص گذاری خودکار و استخراج کلمات کلیدی برای بازیابی اطلاعات و خوشه‌بندی متون". فصلنامه علمی-پژوهشی پردازش علائم و داده‌ها، جلد ۱۳، شماره ۱، صفحه ۷۸-۱۰۰، ۱۳۹۵.

[32] Farhad Rad, Hamid Parvin, Atoosa dahbashi and Behrooz Minaei. "Improved Clustering Persian Text Based on Keyword Using Linguistic and Thesaurus Knowledge", Signal and Data Processing, Vol. 13, No. 1, P.P. 78-100, 2016.

**جواد داودی مقدم** فارغ‌التحصیل کارشناسی نرم‌افزار رایانه از دانشگاه صنعتی بیرجند در سال ۱۳۹۱ و کارشناسی ارشد هوش مصنوعی از دانشگاه صنعتی خواجه‌نصیرالدین طوسی در سال ۱۳۹۳ است. ایشان از مقطع کارشناسی ارشد فعالیت‌های پژوهشی خود را بر روی پردازش متون فارسی آغاز کردند و با بهره‌گیری از روش‌های هوش مصنوعی و داده‌کاوی ابزارهای مختلفی را برای پردازش متون فارسی به‌خصوص در حوزه خبر طراحی و پیاده‌سازی کرده‌اند که مورد بهره‌برداری صنعت نیز قرار گرفته است. همچنین تاکنون دو مقاله کاربردی از ایشان در حوزه پردازش متون فارسی در کنفرانس‌های بین‌المللی داخلی ارائه و چاپ شده است. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: داده‌کاوی و پردازش متن، طراحی و پیاده‌سازی ابزارهای پردازش زبان فارسی، فهم جستار در موتور جستجو، خصوصی‌سازی نتایج موتور جستجو بر

اساس زمینه کاربر، گراف دانشی، سامانه‌های هوشمند و عامل‌های یادگیر.

رایانامه ایشان عبارت است از:

**j.davoudi@chmail.ir**

**علی احمدی** دارای مدرک کارشناسی برق از دانشگاه صنعتی امیرکبیر در سال ۱۳۶۹ و کارشناسی ارشد و دکترای هوش مصنوعی از دانشگاه ایالتی اوزاکای ژاپن در سال‌های ۲۰۰۱ و ۲۰۰۴ میلادی است. ایشان پس از پایان دوره دکترا به‌مدت سه سال در مرکز پژوهش‌های سامانه‌های نانو در دانشگاه هیروشیمای ژاپن مشغول پژوهش در زمینه "طراحی و پیاده‌سازی سخت‌افزاری مدل‌های یادگیر هوشمند" بوده‌اند که منجر به چاپ مقالات علمی در این زمینه شده است. از سال ۱۳۸۶ تاکنون ایشان به‌عنوان استادیار دانشکده کامپیوتر دانشگاه صنعتی خواجه‌نصیرالدین طوسی و از سال ۱۳۸۹ تاکنون به‌عنوان رئیس پژوهشکده فناوری اطلاعات این دانشگاه مشغول به فعالیت هستند. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: وب‌کاوی و موتورهای جستجوی متنی و تصویری، طراحی محیط‌های واقعیت مجازی و تعاملی، طراحی ترکیبی سخت‌افزاری-نرم‌افزاری مدل‌های یادگیر، محاسبات نرم.

رایانامه ایشان عبارت است از:

**ahmadi@eetd.kntu.ac.ir**