استخراج شبکه کلمات هم‌رخداد جهت بهبود پیشنهاددهنده‌ی جستجو در پایگاه اطلاعات علمی ایران "سامانه گنج"

فرزانه بیات[۱]، سمیه فتاحی[*۲]

۱- دانشجوی کارشناسی ارشد گروه کامپیوتر، دانشکده فنی مهندسی، دانشگاه آزاد اسلامی واحد تهران شمال ، تهران، ایران

۲- استادیار پژوهشکده فناوری اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران، ایران

خلاصه

امروزه، با حجم فراوان اطلاعات در وب، یافتن اطلاعات مربوط به یک موضوع خاص، دشوار است. استفاده از سیستم‌های جستجو و جستجوی کلید واژه برای جمع‌آوری سند و داده‌های مرتبط استفاده می‌شود. یک نمونه جستجوی جدید، جستجوی فوری است که در آن هر اعمال کلید توسط کاربر یک درخواست جدید برای سرور ایجاد می‌کند. از کاربردهای جستجوی فوری در پایگاه داده‌های علمی می‌توان نام برد که حجم و تنوع اطلاعات آن‌ها هر روز گسترش می‌یابد و به عنوان یک مخزن قابل اعتماد از آخرین مطالعات انجام شده توسط محققان در زمینه‌های مختلف علم مورد استفاده قرار می‌گیرد. در این میان پایگاه اطلاعات علمی ایران (گنج) با صدها هزار رکورد که بیشتر آن‌ها پایان‌نامه و رساله هستند دارای هزاران مراجعه و ده‌ها هزار جست‌وجو در روز است. در این سامانه با ورود و جستجوی کاربران، لاگ‌هایی ایجاد می‌شود که شامل اطلاعاتی مانند عبارت جستجو شده هستند. در این پژوهش با بررسی این لاگ‌ها، هم‌رخدادی کلمات در جستجوها استخراج شده و شبکه‌ی آن‌ها نمایش داده می‌شود. این شبکه جهت کمک به جستجوی فوری مورد استفاده قرار می‌گیرد. جهت انجام این پژوهش دستورالعمل‌های استخراج متن بر روی داده‌های لاگ اعمال شده و عبارات کلیدی استخراج شدند. یک مجموعه داده از کلمات کلیدی به عنوان گره و فرکانس تکرار آن‌ها با کلمات متناظرشان به عنوان یال تشکیل گردید و شبکه متناظر آن‌ها ایجاد شد. نتایج نشان می‌دهد که این شبکه ترتیب رخداد کلمات در جستجوی کاربران را نمایش می‌دهد که می‌تواند جهت تقویت موتور جستجو و پیشنهاد کلمات کلیدی به کاربران مورد استفاده قرار گیرد.

کلمات کلیدی: کلمات هم‌رخداد، لاگ کاوی، پایگاه اطلاعات علمی ایران(گنج)، موتور جستجو

۱. مقدمه

حجم فراوان و روبه رشد اطلاعات بر روی اینترنت، مشکلاتی برای کاربران به وجود آورده است. زمانی که کاربر به منظور تامین خواسته‌ای به اینترنت متصل می‌شود با حجم عظیم اطلاعات مرتبط با نیازش مواجه می‌شود. انتخاب بهترین مورد از میان موارد موجود، مشکلی است که اکثر کاربران آن را تجربه کرده‌اند[۱]. به عنوان مثال، در سایت‌های تجارت الکترونیک، کاربران باید مدت زمان زیادی را برای پیدا کردن محصولات مورد نظر خود صرف کنند که در نهایت هم ممکن است به نتیجه مطلوب خود نرسند و سایت را برای همیشه رها کنند[۲]. یک راه غلبه بر این مشکلات استفاده از سیستم‌های

[*] **Corresponding author: استادیار پژوهشکده فناوری اطلاعات، پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)، تهران،ایران**
Email: fatahi@irandoc.ac.ir

پیشنهاددهنده است که در موتورهای جستجو به کار می‌روند. موتورهای جستجو، فعالیت‌های کاربر را در قالب لاگ‌های پرس‌وجو ذخیره می‌کنند[۳]. که این داده‌های لاگ به عنوان اطلاعات کلیک شناخته می‌شوند. اکثر موتورهای جستجو که این ذخیره را انجام می‌دهند از جستجوی فوری پشتیبانی می‌کنند. هدف سیستم‌های جستجوی فوری ارائه پیشنهاد کلمات مشابه با جستجوی کاربر می‌باشد. جستجوی فوری نه تنها می‌تواند زمان جستجو را کوتاه کند، بلکه به کاربران کمک می‌کند زمانی که اطلاعات جزئی برای جستجو دارند پاسخ خود را پیدا کنند. مقایسه‌ی سیستم جستجوی فوری با جستجوی مرسوم نشان می‌دهد که جستجوی فوری می‌تواند با ارسال پاسخ‌های مربوطه قبل از اینکه کاربر تایپ کردن پرس‌وجو را کامل کند، نیاز به تایپ کامل کلمه جستجو را کم کند. یکی از کاربردهای جستجوی فوری، در پایگاه داده‌های علمی است که حجم و تنوع اطلاعات آن‌ها هر روز گسترش می‌یابد. استفاده از این پایگاه‌های داده به عنوان یک مخزن قابل اعتماد از آخرین مطالعات انجام شده توسط محققان در زمینه‌های مختلف علم افزایش یافته است. در این میان یکی از سیستم‌هایی که در ایران امکان جستجو برای پژوهشگران را فراهم می‌کند، سامانه گنج پژوهشگاه علوم و فناوری اطلاعات ایران است. سامانه گنج (گنجینه علمی ایرانیان) پژوهشگاه علوم و فناوری اطلاعات ایران با برخورداری از صدها هزار مدرک علمی، امکان جستجو برای پژوهشگران را در پایان‌نامه‌ها، نشریات علمی داخلی، مقالات، همایش‌ها، طرح‌های پژوهشی وگزارش‌های دولتی فراهم می‌کند. این پایگاه داده که کاربران آن به طور عمده محققان دانشگاهی داخل و گاهی خارج از کشور، دانشجویان کارشناسی ارشد و دکتری هستند، بیش از ۹۰۰۰ بازدید کننده آنلاین در هر روز دارد و روزانه بیش از ۵۰۰۰۰ پرس‌وجو در این سامانه توسط کاربران حقیقی صورت می‌گیرد[۴]. با بررسی داده‌های لاگ جستجویی که توسط کاربران در سامانه گنج انجام شده است، می‌توان شبکه‌ای از کلمات هم‌رخداد استخراج کرد که در جهت بهبود این سامانه مورد استفاده قرار گیرد.

۲. پیشینه پژوهش

امروزه تحقیقات بسیاری در زمینه تعامل انسان و کامپیوتر جهت افزایش رضایت کاربران در زمینه استفاده از کامپیوتر انجام شده است [۱۴] [۱۵] [۱۶]. یکی از این حوزه‌ها استفاده از موتورهای جستجو است. استخراج مفاهیم اصلی از مقدار زیادی از اطلاعات علمی در میان پژوهش‌ها بیش از پیش مورد توجه واقع شده است[۵]. برای رسیدن به این هدف پروژه‌های زیادی انجام گرفته است. علیرغم تلاش‌های بسیار محققان، تحقیقات بسیار کمی در زمینه کشف کلمات هم‌رخداد از لاگ صورت گرفته است. در هر تجربه، محققان از دیدگاه‌های مختلف به این هدف نگاه می‌کنند که در این بخش به برخی از مهم‌ترین آن‌ها پرداخته می‌شود:

با افزایش چالش‌ها در زمینه استخراج متن فارسی در طی سالیان متمادی، کار بر روی پردازش متن‌های فارسی و توسعه مجموعه داده‌های جمع‌بندی فارسی مورد توجه بسیاری از محققان قرار گرفت. در سال ۲۰۰۸ بیدگلی و ایرانپور[۶] با محاسبه رخداد همزمان پسین و پیشین، عبارت‌های کلیدی بیش از یک واژه‌ای را استخراج کردند. یکی از مزیت‌های این روش سرعت بالا نسبت به روش TFIDF می‌باشد. استفاده از روش رخداد همزمان بهبود خوبی را نشان می‌دهد و واژه‌های کلیدی با معنی پیشنهاد می‌کند.

یکی از مراحل جمع‌بندی متن فارسی، مهار و از بین بردن کلمات متداول است. در سال ۲۰۰۹ برنجی‌کوب و همکاران[۷]، از ترکیب تکنیک‌های موجود در شبکه کلمات برای ایجاد یک بانک اطلاعاتی فارسی با استفاده از فرهنگ لغت دهخدا استفاده کرده‌اند و الگوریتم‌های جدیدی برای ایجاد کلمات در اسناد فارسی نیز ارائه شده است. روش پیشنهادی آن‌ها در مقایسه با نتایج سایر الگوریتم‌های متون فارسی ریشه کلمات موجود را با دقت بیشتری استخراج می‌کند.

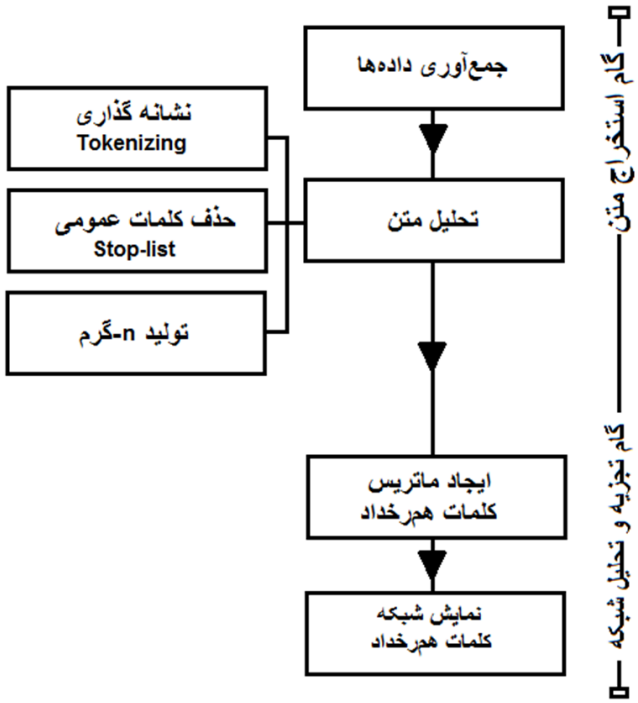
در سال ۲۰۱۱ خوزانی و همکاران[۸] روشی در مورد استخراج کلمات کلیدی یک کلمه و دو کلمه فارسی مورد بررسی قرار دادند. مجموعه‌ای شامل بیش از ۱۰۰۰۰۰ سند را انتخاب کردند که از این میان ۴۸ سند با تعداد کل ۱۰۰۰ تا ۲۵۰۰ کلمه در هر سند مورد تجزیه و تحلیل قرار گرفت. مقایسه‌های انجام شده بین یافته‌های این روش‌ها و دیگر روش‌ها ثابت کرده است که این روش کلمات کلیدی متون را با دقت و سرعت بسیار بیشتری استخراج می‌کند تا مفاهیم اصلی را نشان دهد.
در سال ۲۰۱۲ شعبان و همکاران[۹] از روش تحلیل هم‌رخدادی کلمات استفاده کرده‌اند و نقشه مفهومی دانش با استفاده از بررسی ۳۰۰ مقاله منتشر شده در حوزه توسعه نوآوری در مناطق از سال ۱۹۹۰ تا ۲۰۱۲ در پایگاه‌های اطلاعاتی اسکوپوس و سیج* ترسیم کردند. براساس نتایج این روش، جریان‌های غالب در حوزه توسعه نوآوری در مناطق در قالب سه جریان اصلی ارائه شده است. ایجاد چنین دسته‌بندی به سیاستگذاران در امر شناخت بهتر از جریان‌های غالب در توسعه نوآوری در مناطق جهت اتخاذ سیاست‌های ثمربخش‌تر کمک می‌نماید.
در سال ۲۰۱۲ جلالی‌منش و همکاران[۵] روش جدیدی را برای استخراج مفاهیم اصلی از مجموعه متنی معرفی می‌کنند که بر مبنای تجزیه و تحلیل داده‌های متنی و شبکه اجتماعی است. روش را با یک متن متشکل از ۶۵۰ عنوان پایان‌نامه مهندسی صنایع از مخزن ایرانداک آزمایش کرده‌اند. این مطالعه نشان می‌دهد که تجزیه و تحلیل شبکه یک راه موثر برای استخراج مفاهیم اصلی از متن است و نمایش شبکه‌ی اصطلاحات نمای جامع از مخازن متن ارائه می‌دهد.
در سال ۲۰۱۳ دانش و همکاران[۱۰] روشی برای بازیابی اسناد فارسی با استفاده از تکنیک فهرست‌بندی و توزیع N-gram ارائه کردند. فهرست پیشنهادی روشی برای نمایش سؤالات پاسخگو موثرتر است که کیفیت بازیابی اطلاعات را به میزان قابل توجهی افزایش می‌دهد و بازیابی بهینه‌تر را در اسناد فارسی به دست می‌آورد. اما سرعت ایندکس N-گرم کم است. برای حل این مشکل یک مکانیسم نمایه سازی N-گرم توزیع شده برای سیستم‌های بزرگ فارسی زبان طراحی کرده‌اند. با سایر روش‌های موجود در این زمینه، کیفیت اسناد بازیابی شده و همچنین سرعت بازیابی اطلاعات را بهبود می‌بخشد.
در سال ۲۰۱۴ صادقی و همکاران[۱۱] روشی برای ایجاد خودکار لیست توقف برای یک سیستم بازیابی اطلاعات متنی فارسی شرح داده‌اند که می‌تواند برای پردازش زبان طبیعی و به ویژه برای اهداف بازیابی اطلاعات استفاده شود. نتایج نشان می‌دهد که کلمات توقف، مشخص شده با اصطلاحات فرکانس بالا در مجموعه، روشی بسیار مؤثر برای بدست آوردن لیست کلمات توقف است، زیرا بین این لیست و لیست نهایی به دست آمده بهترین شباهت و فاصله کمتری وجود دارد.
در سال ۲۰۱۶ صلواتی و همکاران[۱۲] روش‌های موجود در اندازه‌گیری شباهت بین صفحات براساس معیار رخداد کلمات با استفاده از رپیدماینر، جفت کلمات از صفحات استخراج کردند و با استفاده از موتور جستجوی گوگل مورد بررسی قرار دادند. تعداد صفحات با استفاده از روش‌های ضریب همپوشانی، ضریب شباهت کسینوسی، ضریب سیمپسون و ضریب PMI مقایسه گردید. نتایج بدست آمده، بیانگر میزان شباهت معنایی بهتر بین صفحات در صورت ارتباط معنایی در کلمات رخداده شده می‌باشند.
در سال ۲۰۱۹ کرمانی و همکاران[۱۳] پس از انجام پیش پردازش و بررسی تکنیک‌های پیشرفته در خلاصه‌کردن و تجزیه و تحلیل مقالات برتر اخبار فارسی از جمله همشهری، ایسنا و تابناک، یک رویکرد خلاصه‌ای استخراج کردند که ترکیبی از خصوصیات متن آماری و مفهومی است و همچنین ساختار این سند ارائه شده و نشان داده شده است تا بطور خودکار خلاصه فارسی باشد. امتیازدهی از ویژگی‌های برجسته بر اساس عنوان مقاله، اسم‌های مناسب، ضمایر، طول جمله، کلمات کلیدی، کلمات موضوع، موقعیت جمله، کلمات انگلیسی و نقل قول‌ها است. این سیستم از پنج بخش اصلی تشکیل شده است که شامل پیش پردازش متن فارسی، تولید ویژگی‌ها، امتیازدهی، پردازش پس از حذف اطلاعات اضافی و در نهایت نتیجه‌گیری

* sage

خلاصه تولید می‌شود. این روش با استفاده از مجموعه داده‌های Pasokh مورد بررسی قرار گرفته است و نتایج مقایسه شده نشان می‌دهد که عملکرد بالاتری دارد.
بطور کلی می‌توان نتیجه گرفت که در سال‌های اخیر، تکنیک‌های زیادی در این زمینه ارائه شده است. اما در زمینه لاگ کاوی جهت بهبود سیستم‌های پیشنهاددهنده تکنیک‌های چندانی یافت نشده است. همچنین نیاز بهبود پایگاه‌های علمی مشابه سامانه گنج به شدت احساس می‌شود که تحقیقی در این زمینه انجام نشده است. بنابراین در این مطالعه یک روش استخراج شبکه کلمات هم‌رخداد جهت بهبود سامانه گنج ارائه شده است.

۳. روش شناسی پژوهش

روند کلی پژوهش در شکل ۱ آمده است. ما این فرآیند را به دو مرحله اصلی تقسیم می‌کنیم که عبارتند از: استخراج متن و تجزیه و تحلیل شبکه. در ابتدای کار تحلیل داده‌های لاگ صورت گرفته است و سپس نمایش کلمات هم‌رخداد به صورت گراف قابل درک نمایش داده شده است که نتیجه‌ی این روند جهت بهبود عملکرد موتورهای جستجو در قسمت پیشنهاددهنده‌ی سیستم قابل استفاده است.



شکل۱- شمای کلی پژوهش

در این پژوهش داده‌های مورد استفاده از سامانه جستجوی گنج و از بخش لاگ‌های کاربران استخراج شده است. این کاربران با عضویت در سامانه و یا به صورت مهمان وارد سامانه گنج شده‌اند. در زمانی که کاربران از موتور جستجوی این سامانه استفاده می‌کنند، اطلاعات آنان ذخیره می‌شود. این اطلاعات شامل پرس‌و جو، شناسه کاربر، تاریخ و زمان جستجو هستند که در این پژوهش پرس‌و جوها مورد مطالعه قرار گرفته‌اند. نمونه‌ای از این داده‌های لاگ در شکل ۲ آمده است.

	User_id	session_id	ip_addres	url	info	query	created_at
2	43900318	3663555a7	192.168.1	/api/v1/	{\result_count\"=>36	\"search_title\"=>\"\\\"روانشناسی فروش\\\"\"}	2019-05-26 00:20:15
3	43900357	3663555a7	192.168.1	/api/v1/	{\result_count\"=>1	\"search_title\"=>\"(عنوان(روانشناسی) و عنوان(فروش\"}	2019-05-26 00:20:16
4	43902563	294a6e826(	192.168.1	/api/v1/	{\result_count\"=>57	\"search_title\"=>\"\\\"روانشناسی بالینی\\\" و \\\"روانشناسی عمومی و میان رشته ای\\\"\"}	2019-05-26 00:44:03
5	43902962	8dba24b50(	192.168.1	/api/v1/	{\result_count\"=>0	\"search_title\"=>\"\\\"1392\\\" و \\\"1391\\\" و \\\"1390\\\" و \\\"1377\"	2019-05-26 00:47:12
6	43904288	31d3e331cd	192.168.1	/api/v1/	{\result_count\"=>356	\"search_title\"=>\"\\\"پایان نامه روانشناسی\\\"\"}	2019-05-26 01:00:14
7	43904630	cae0337d71	192.168.1	/api/v1/	{\result_count\"=>228	\"search_title\"=>\"\\\"روانشناسی رنگ\\\"\"}	2019-05-26 01:03:01
8	43904954	dfddc8d5b6	192.168.1	/api/v1/	{\result_count\"=>4396	\"search_title\"=>\"\\\"روانشناسی\\\"\"}	2019-05-26 01:09:33
9	43905723	cae0337d71	192.168.1	/api/v1/	{\result_count\"=>35	\"search_title\"=>\"\\\"روانشناسی رنگ در ایران\\\"\"}	2019-05-26 01:21:12
10	43913846	309950e8a5	192.168.1	/api/v1/	{\result_count\"=>4	\"search_title\"=>\"\\\"شهادت کیفری روانشناسی\\\"\"}	2019-05-26 02:46:56
11	43915237	5bcb5398df	192.168.1	/api/v1/	{\result_count\"=>11	\"search_title\"=>\"\\\"تعصّب دینی\\\" و \\\"دینداری\\\" و \\\"دینداری\\\" و \\\"	2019-05-26 03:02:28
12	43915669	f834d76016	192.168.1	/api/v1/	{\result_count\"=>11	\"search_title\"=>\"\\\"شهادت در روانشناسی\\\"\"}	2019-05-26 03:06:05
13	43918998	62602d0017	192.168.1	/api/v1/	{\result_count\"=>84	\"search_title\"=>\"\\\"دانشکده فنی و مهندسی\\\" و \\\"کده علوم تربیتی و روانشناسی\\\" و	2019-05-26 03:40:53
14	43920913	63964ab84(	192.168.1	/api/v1/	{\result_count\"=>74	\"search_title\"=>\"\\\"مدل ساختاری روانشناسی\\\"\"}	2019-05-26 03:59:33
15	43921261	c607faef02	192.168.1	/api/v1/	{\result_count\"=>0	\"search_title\"=>\"\\\"انهی از منکر بدحجابان با رویکرد فقهی، تربیتی و روانشناسی\\\"	2019-05-26 04:03:17
16	43922204	d38cd2bf0e	192.168.1	/api/v1/	{\result_count\"=>0	\"search_title\"=>\"\\\"مقاله روانشناسی نابینا\\\"\"}	2019-05-26 04:13:42
17	43922708	bb4edbb81c	192.168.1	/api/v1/	{\result_count\"=>0	\"search_title\"=>\"\\\"خلع بدن فراروانشناسی\\\" در عنوان\"}	2019-05-26 04:20:24
18	43925633	383ddac884	192.168.1	/api/v1/	{\result_count\"=>838	\"search_title\"=>\"\\\"روانشناسی عمومی و میان رشته ای\\\" و \\\"روانشناسی بالینی\"	2019-05-26 04:44:56
19	43925844	2b6955a56(	192.168.1	/api/v1/	{\result_count\"=>75	\"search_title\"=>\"\\\"دکتری تخصصی\\\" محدود به مقطع برابر \\\"ضایت زناشویی\\\"\"	2019-05-26 04:48:14
20	43926353	2b6955a56(	192.168.1	/api/v1/	{\result_count\"=>160	\"search_title\"=>\"\\\"دانشکده علوم تربیتی و روانشناسی\\\" و \\\"ی و علوم تربیتی\\\"\"	2019-05-26 04:55:02
21	43926552	2b6955a56(	192.168.1	/api/v1/	{\result_count\"=>160	\"search_title\"=>\"\\\"دانشکده علوم تربیتی و روانشناسی\\\" و \\\"ی و علوم تربیتی\\\"\"	2019-05-26 04:56:38

شکل۲- نمونه‌ای از داده‌های لاگ

بخش استخراج متن شامل نشانه گذاری tokenizing و چک کردن stop-list است. متن پیوسته باید به کلمات، عبارات، نمادها و سایر عناصرکه به token شناخته می‌شوند علامت گذاری شود. پس از آن کلمات استخراج شده با stop-list برای بررسی کلمات غیر ضروری مورد بررسی قرار می‌گیرند. stop-list (یا "کلمات عمومی") لیستی از کلمات بدون اطلاعات هستند. نمونه‌ای از این کلمات در شکل ۳ موجود است.

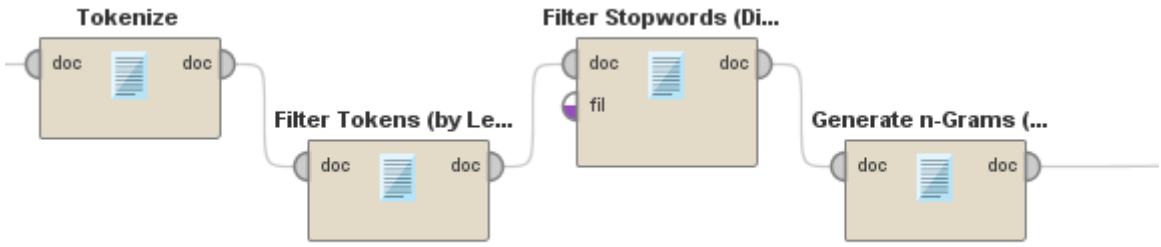
کردن	درباره	همچنین	گرفته	صورت	چه	آن
کنیم	بعد	کردند	هایی	یکی	همچنین	یک
کنید	نشان	مختلف	تواند	هستند	کردند	شود
سازی	حتی	گیرد	اول	است	داده	شده
سوم	اینکه	شما	بار	من	بوده	خود
کنم	ولی	گفته	چند	دهد	دارند	ها
بلکه	توسط	آنان	جدید	هزار	همین	کرد
زیر	چنین	نه	بیش	نیست	سوی	شد
توانند	برخی	طور	شدن	داد	شوند	ای
ضمن	برخی	گرفت	کردن	داشته	بیشتر	تا
فقط	دیروز	دهند	کنیم	داشت	بسیار	کند
دوم	گذاری	درباره	روی	چه	روی	بر

شکل۳- کلمات عمومی

پس از آن N-gram ها شناسایی می‌شوند. N-gram یک دنبالهٔ همسایه N موردی از یک دنباله‌ی متنی است. N-gram های شناخته شده مورد بررسی قرار می‌گیرند و در پایان این مرحله مجموعه‌ای از n-gram های استخراج شده داریم. مرحله دوم روش ما ترسیم و تحلیل شبکه است و متن به عنوان یک شبکه نمایش داده می‌شود. در این بخش، یک شبکه وزن‌دار

تولید شده است سپس شبکه با بسته‌های نرم افزاری SNA* تجسم می‌شود. بسته‌های نرم‌افزاری زیادی با ویژگی‌های مختلف از جمله Gephi وجود دارد که می‌تواند برای این منظور استفاده شود. در آخر لیستی از کلمات هم‌رخداد خواهیم داشت.

برای فاز متن کاوی نرم افزار RapidMiner 9.5 (یک بسته open source) مورد استفاده قرار گرفته است.



شکل۴- نرم افزار رپیدماینر

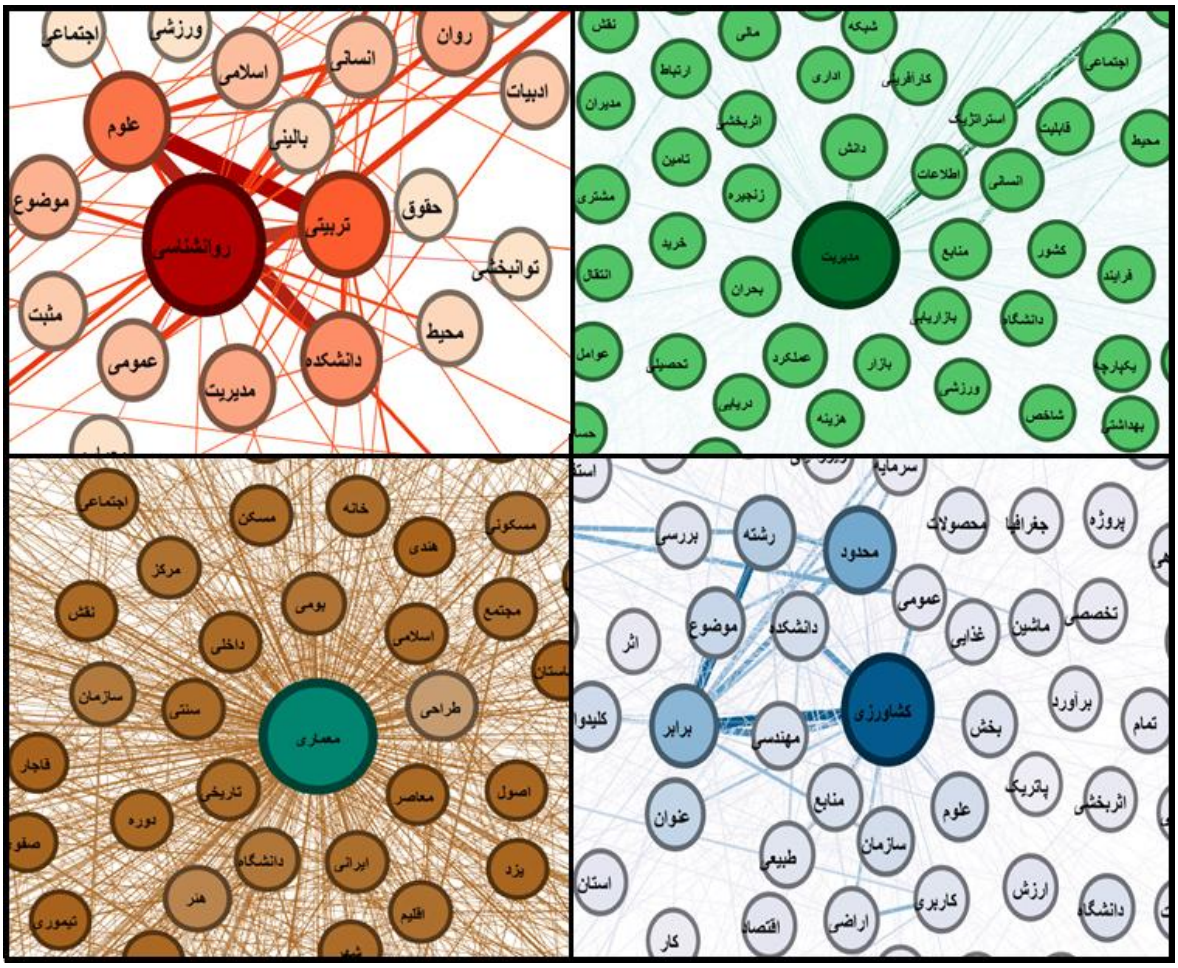
پس از حذف کلمات عمومی و کشف عبارت‌های کلیدی، به بررسی تعداد کلمات در هر عبارت پرداخته شده است. اصطلاح n-گرم به عنوان یک سری نشانه‌های متوالی از طول n تعریف می‌شود و شامل کلیه سری علائم متوالی از طول n است. حداکثر طول n-Grams عدد صحیح است. در این پژوهش برای کشف عبارات کلیدی به صورت پیش فرض عباراتی که دارای حداکثر سه کلمه بودند انتخاب شده و مورد مطالعه قرار گرفتند. به این منظور در تولید n-گرم‌ها در این عملگر $n=3$ در نظر گرفته شده است. این مراحل به حدی تکرار می‌شوند تا در نهایت لیستی از کلمات کلیدی فاقد کلمات عمومی که با علامت "-" از هم جدا شده‌اند در نتایج ظاهر می‌شوند. پس از متن‌کاوی و مشخص شدن کلمات کلیدی در این بخش به محاسبه تعداد وقوع کلمات پرداخته شده است. مطابق با روش، یک ماتریس کلمات هم‌رخداد ایجاد گردید. نتایج نهایی از گام استخراج متن به دست آمده و تمام سطرها به صورت لیست در متغیر قرار می‌گیرد. هر سطر لیست به دو قسمت تقسیم شده است قسمت اول همان اصطلاحات هستند و قسمت دوم تعداد دفعات مشاهده می‌باشد. بخش کلماتی که در یک نشست وجود دارند از هم جدا در نظر گرفته شده‌اند تا در مرحله‌ی نمایش شبکه هرکدام به صورت یک گره نمایش داده شوند. تعداد تکرار این کلمات محاسبه شده و بعد از آن‌ها هر کلمه‌ای استفاده شود نیز محاسبه می‌شود. به این معنی که ترتیب استفاده‌ی کلمات در یک عبارت نیز در نظر گرفته می‌شود. تمامی اصطلاحات به عنوان گره و تعداد دفعات مشاهده به عنوان یال ذخیره می‌شوند و نتیجه شامل مقادیر بالایی گره و یال است. در این مرحله اطلاعات نهایی ایجاد شده است و در قالب GraphML (XML) جهت نمایش درنظر گرفته می‌شود.

گفی یک ابزار تجزیه و تحلیل است که گراف شبکه را نمایش می‌دهد و تجزیه و تحلیل می‌کند. همچنین دارای برخی ویژگی‌های محاسبه اقدامات شبکه و خوشه‌بندی شبکه است. گفی، انتخاب روش‌های مختلف، تنظیم گزینه‌های خاص و مشاهده نتایج را ساده می‌کند. به همین منظور در این پژوهش مورد استفاده قرار گرفته است.

۴. تجزیه و تحلیل یافته‌ها

* Social Network Analysis

در نرم افزار گفی، پس از وارد کردن نتیجه‌ی مراحل قبل که ماتریسی از گره‌ها و یال‌های آن‌ها بوده است، شبکه‌ی داده‌ها ترسیم شده است. این شبکه دارای تعداد زیادی گره و یال است و جهت مشاهده و درک بهتر پس از اولین اجرای گفی، کلمات کلیدی مهم‌تر جدا شدند و سپس به ترسیم شبکه‌ی آن‌ها به صورت جداگانه پرداخته شد تا نتایج مشخص و قابل درک باشند. سایز گره‌ها نشان دهنده تعداد رخداد برای هر اصطلاح است و ضخامت یال نشان دهنده تعداد همپوشانی برای هر دو اصطلاح است. بر اساس این رقم، واژه‌هایی نظیر روانشناسی، مدیریت، معماری و کشاورزی واژه‌هایی هستند که بارها استفاده شده‌اند.



شکل۵- گراف‌های متمرکز بر کلمات روانشناسی، مدیریت، معماری و کشاورزی

همان‌طور که در شکل ۵ مشهود است کلمات پر استفاده یکی پس از دیگری با یال به هم متصل هستند و هم‌رخدادی کلمات قابل مشاهده است. با این حال جدولی به این منظور رسم شده است. برای مثال یکی از پرتکرارترین کلماتی که در لاگ‌ها استفاده شده بود کلمه‌ی "مدیریت" است که گره‌ی بزرگتری دارد و یال‌های متعددی به آن متصل است. با حذف یال‌ها و گره‌های نسبتا دور در شبکه به شبکه‌ی مفهوم‌تر و مهم‌تری دست پیدا کردیم. همان‌طور که مشاهده می‌شود کلمه‌ی "مدیریت" به همراه کلمات "دانش" ، "شبکه" و "بحران" و کلمه‌ی دانش نیز به همراه کلمه‌ی سلامت بیشتر استفاده شده است. به این صورت که عبارت "مدیریت دانش سلامت" در یک نشست به وجود آمده است و این نشست

بیشتر از نشست‌های مشابه تکرار شده است. برای هر کدام از سه کلمه یک گره اختصاص داده شده است و به وسیله یال به هم متصل هستند. در این شبکه همان‌طور که قابل مشاهده است عبارات پر تکرار را در جدول۱ نیز نمایش داده‌ایم:

جدول ۱- کلمات هم‌رخداد با کلمه‌ی "مدیریت"

مدیریت	دانش	
مدیریت	دانش	کارآفرینی
مدیریت	دانش	سلامت
مدیریت	شبکه	اطلاعات
مدیریت	شبکه	رایانه ای
مدیریت	بحران	دریایی
مدیریت	عملکرد	هزینه
مدیریت	ورزشی	دانشگاه
مدیریت	بازاریابی	
مدیریت	بازاریابی	ورزشی

۵. نتیجه‌گیری

در این مطالعه از طریق داده‌های لاگ جستجویی که توسط کاربران انجام شده است، شبکه‌ای از کلمات هم‌رخداد استخراج شد تا در جهت بهبود موتور جستجوی سامانه گنج استفاده گردد. جهت انجام پژوهش، داده‌های لاگ سامانه گنج در مدت ۳ ماه جمع‌آوری شد و بخش‌هایی از آن به عنوان نمونه استفاده شدند .در مرحله اول دستورالعمل‌های استخراج متن بر روی این داده‌ها اعمال شد و عبارات کلیدی استخراج شدند. یک مجموعه داده از کلمات کلیدی به عنوان گره و فرکانس تکرار آنها با کلمات متناظرشان به عنوان یال تشکیل گردید. این بدان معنی است که وقتی دو کلمه یا عبارت در یک رکورد از لاگ‌های جستجو توسط کاربر ظاهر می‌شوند، به یکدیگر متصل می‌شوند. سپس یک شبکه تولید شد که هر گره نمایانگر یک کلمه است و یال‌ها کلمات را به یکدیگر پیوند می‌دهند، به طوری که قدرت یک یال نشان دهنده چگونگی استفاده دو کلمه در جستجوهای کاربر کنار یکدیگر است. قدرت یک یال براساس فرکانس تکرار آن در نمونه‌های لاگ تعریف می‌شود. یعنی اگر دو کلمه "انرژی" و " خورشیدی" ۳ بار توسط کاربر کنار یکدیگر جستجو شده باشند و کلمات " انرژی" و " الکتریکی" ۵ بار کنار یکدیگر جستجو شده باشند، قدرت یال بین گره " انرژی" و " الکتریکی" ۲ بار قوی‌تر از قدرت یال بین دو گره "انرژی" و " خورشیدی" است. همچنین براساس تعداد رخداد هر کلمه گره‌ی متناظر آن بزرگ‌تر نمایش داده شده است. سپس این شبکه با استفاده از تکنیک‌های SNA تحلیل می‌شود. در این پژوهش داده‌های لاگ جستجوی کاربران در سامانه گنج مورد بررسی قرار گرفتند. به منظور استخراج کلمات کلیدی، فرآیند متن کاوی بر روی داده‌ها اعمال گردید و پس از ایجاد ماتریس کلمات هم‌رخداد، نتایج به صورت شبکه نمایش داده شد. در آخر برخی از گراف‌های مهم‌تر و لیست‌های کلمات هم‌رخداد که در شبکه‌های ترسیم شده موجود است نمایش داده شد. بررسی لاگ‌های پرس‌وجو به پژوهشگران کمک می‌کند تا فرآیند جستجو را بهبود دهند و با مشخص شدن کلمات هم‌رخداد می‌توان کاربران را در یافتن نتایج مناسب پرس‌وجو راهنمایی نمود. به این صورت که هنگامی که کاربر کلمه‌ای را تایپ می‌کند براساس داده‌های گذشته‌ی

سیستم کلمه‌های بعدی که بیشتر جستجو شده‌اند را پیشنهاد بدهد. با توجه به رشد سریع تکنولوژی و تعداد روزافزون کاربران این سامانه می‌توان در آینده براساس جنسیت کاربران، رشته تحصیلی و یا از نظر زمانی جستجوهای سال‌های مختلف که جامعه با سیاست‌های مختلفی همراه بوده است را مورد بررسی قرار داد.

۶. مراجع

۱. نوروزی، س. ، م. مقیمی، و ر. رافع، ارائه یک سیستم پیشنهادگر برای پیش بینی رفتار کاربران وب، در هفتمین کنفرانس بین المللی فناوری اطلاعات و دانش. ۱۳۹۴.

۲. صیاد، م. و ع.ا. خنجری میانه، یک سیستم پیشنهاددهنده وب براساس ترکیب داده کاوش استفاده وب وجلسه فعال کاربر، در اولین همایش ملی رویکردهای نوین در مهندسی کامپیوتر و بازیابی اطلاعات. ۱۳۹۲، دانشگاه آزاد اسلامی واحد رودسر و املش.

3. Cetindil, I., et al. *Analysis of Instant Search Query Logs*. in *WebDB*. 2012. Citeseer.

4. Rabiei, M., S.-M. Hosseini-Motlagh, and A.J.S. Haeri, *Using text mining techniques for identifying research gaps and priorities: a case study of the environmental science in Iran.* 2017. **110**(2): p. 815-842.

5. Jalalimanesh, A. *Knowledge discovery in scientific databases using text mining and social network analysis*. in *2012 IEEE Conference on Control, Systems & Industrial Informatics*. 2012. IEEE.

٦. ابرانپور، م. و ب. مینایی بیدگلی، یک روش جدید برای استخراج کلمات و عبارات کلیدی تک سنده فارسی با استفاده از تعیین حدود جمله، در دومین کنگره مشترک سیستم های فازی و سیستم های هوشمند. ۱۳۸۷، دانشگاه فردوسی مشهد.

7. Berenjkoob, M., et al. *A method for stemming and eliminating common words for Persian text summarization*. in *2009 International Conference on Natural Language Processing and Knowledge Engineering*. 2009. IEEE.

8. Khozani, S.M.H. and H. Bayat. *Specialization of keyword extraction approach to Persian texts*. in *2011 International Conference of Soft Computing and Pattern Recognition (SoCPaR)*. 2011. IEEE.

۹. شعبان، ا. ، et al. ، شناسایی جریان های غالب در حوزه توسعه نوآوری در مناطق با استفاده از روش تحلیل هم رخدادی کلمات. بهبود مدیریت, ۱۳۹۱. سال ششم(۱۷): p. 136-.

10. Danesh, M., et al., *A Distributed N-Gram Indexing System to Optimizing Persian Information Retrieval.* 2013. **5**(2): p. 214.

11. Sadeghi ,M. and J.J.J.o.I.S. Vegas, *Automatic identification of light stop words for Persian information retrieval systems.* 2014. **40**(4): p. 476-487.

۱۲. صلواتی، س. ، ف. احمدی آبکناری، و ر. مجتهدی صفاری، تقویت معیار شباهت معنایی بین صفحات وب بر اساس درجه معناداری ارتباط کلمات، در سومین کنفرانس بین المللی مهندسی دانش بنیان و نوآوری. ۱۳۹۵، دانشگاه پیام نور استان تهران- فرمانداری شهر قدس - انتشارات KBEI.

13. Kermani, F.H. and S. Ghanbari. *Extractive Persian Summarizer for News Websites*. in *2019 5th International Conference on Web Research (ICWR)*. 2019. IEEE.

14. Fatahi, S. (2019). An experimental study on an adaptive e-learning environment based on learner's personality and emotion. *Education and Information Technologies*, 1-17.

15. Fatahi, S. Ershadi, M. (2019). “Assessment of user satisfaction in national search engine of Dissemination System of Theses and Dissertations (GANJ) based on eQual model”, Iranian Journal of Information Processing and Management

16. Fatahi, S., & Moradian, S. (2018). An Empirical Study on the Impact of Using an Adaptive e-Learning Environment Based on Learner's Personality and Emotion. *International Association for Development of the Information Society*.