فصلنامه تخصصی پارک‌های علم و فناوری و مراکز رشد
فصلنامه رشد فناوری، سال بیستم، شماره ۷۷، زمستان ۱۴۰۲
www.roshdefanavari.ir
شاپا: ۵۴۸۶-۱۷۳۵ / شاپای الکترونیکی: ۵۶۶۴-۱۷۳۵

کاربرد بازیابی هوشمند اطلاعات در جستجوی پتنت

حبیبا... اصغری*
پژوهشکده فناوری اطلاعات و ارتباطات
جهاد دانشگاهی، تهران، ایران
habib.asghari@ictrc.ac.ir

آزاده شاکری
دانشکده مهندسی برق و کامپیوتر،
دانشکدگان فنی دانشگاه تهران، تهران، ایران
shakery@ut.ac.ir

تاریخ دریافت: ۱۴۰۲/۰۲/۱۱ تاریخ اصلاحات: ۱۴۰۲/۰۸/۰۷ تاریخ پذیرش: ۱۴۰۲/۰۹/۱۳

چکیده

در ارزیابی درخواست ثبت اختراع، جستجو در فهرست اطلاعات قبلی نقشی بسیار مهم و حائز اهمیت دارد. عموماً این جستجو توسط افراد خبره انجام می‌شود و فرایندی بسیار زمان‌بر است. جستجو از طریق روش‌های مختلف بازیابی هوشمند اطلاعات می‌تواند نقش مؤثری در فرایند بازیابی اطلاعات همسان ایفا نماید. یکی از مهم‌ترین مسائل مرتبط با بازیابی اسناد پتنت، ایجاد یک عبارت پرس و جوی کارآمد به منظور جستجو می‌باشد. از جمله شیوه‌های ساخت عبارت پرس و جو می‌توان به تولید خودکار پرس و جو از تقاضانامه ثبت پتنت اشاره نمود. همچنین در روش‌های دیگر، از تمامی متن سند پتنت به‌عنوان پرس و جو جهت اجرای جستجو استفاده می‌شود. نکته حائز اهمیت آن است که غالباً به دلیل نحوه خاص نگارش اسناد پتنت و ایجاد ابهامات معنایی، گسترش عبارت پرس و جو از اهمیت ویژه‌ای برخوردار است. در این مقاله مجموعه تحقیقات انجام‌شده درخصوص بازیابی اسناد پتنت با کمک الگوریتم‌های بازیابی هوشمند اطلاعات در هر دو زمینه بازیابی تک زبانی و بین زبانی مورد بررسی قرار می‌گیرد. همچنین معیارهای ارزیابی کیفیت بازیابی و شیوه مرتب‌سازی اسناد بررسی می‌گردد.

واژگان کلیدی

بازیابی هوشمند اطلاعات؛ بازیابی متنی؛ جستجوی پتنت؛ گسترش پرس و جو؛ حق ثبت اختراع.

۱- مقدمه

امروزه بخش مهمی از سرمایه شرکت‌های دانش‌بنیان و مؤسسات تحقیقاتی در جهان را دارایی فکری آنها شکل می‌دهد. حقوق مالکیت فکری عبارت از مجموعه حقوق قانونی و قواعدی است که دارایی‌های فکری و فعالیت‌های ذهنی افراد و شرکت‌ها را حفظ و صیانت می‌نماید.

براساس تعریف سازمان جهانی مالکیت فکری[1]، اختراع راه‌حلی است که برای اولین بار و به منظور حل یکی از مشکلات بشر ارائه می‌گردد. حق ثبت اختراع (پتنت) یکی از حقوق مالکیت فکری است که دولت حق انحصاری بهره‌برداری از اختراع را برای مدت معین به مالک آن واگذار کرده و دیگران را از تولید و فروش آن باز می‌دارد [۱]. شرط حمایت قانونی از اختراع آن است که جزئیات آن افشا گردد. دارنده پتنت می‌تواند اجازه بهره‌برداری از حق ثبت اختراع را از طریق انعقاد قرارداد به دیگران واگذار نماید. گردش مالی تجارت مالکیت فکری با استفاده از ابزار پتنت به چند میلیارد دلار در سال بالغ می‌گردد.

در مرحله اولیه ارسال تقاضانامه ثبت پتنت[2]، لازم است که جستجویی دقیق در پایگاه داده‌های پتنت صورت گیرد. هدف اصلی از جستجو و بازیابی، تعیین اصالت و نوآوری در مرحله درخواست ثبت پتنت است. لذا تمامی پتنت‌های ثبت‌شده تا زمان ارائه و تحویل درخواست ثبت باید به دقت مورد جستجو قرار گیرند. این امر از آن‌رو صورت می‌گیرد که مجموعه اطلاعاتی که قبل از ثبت پتنت افشا شده و در اختیار عموم قرار دارد قابل تبدیل به پتنت نمی‌باشد. به دلیل امکان ایجاد مشکلات حقوقی، از دست دادن تنها یک سند مرتبط می‌تواند به دلیل تخلف حقوقی منجر به طرح دعوی و شکایت در دادگاه شود. لذا مسأله بازیابی پتنت را معمولاً به‌عنوان یک مسأله مرتبط با فراخوانی[3] در نظر می‌گیرند که در آن غالباً هدف از جستجو، یافتن نقاط برون نهشته[4] می‌باشد.

جستجوی ثبت اختراع را می‌توان نمونه‌ای خاص و پیچیده از بازیابی اطلاعات در نظر گرفت، که هدف آن یافتن اطلاعات مرتبط، با ماهیت بدون ساختار در مجموعه عظیم داده است. متن پتنت با متن معمولی متفاوت است. جملات مورد استفاده در اسناد ثبت اختراع معمولاً طولانی‌تر از جملات عمومی هستند. در تحقیقی که توسط ایویاما و همکاران انجام شده است طول اسناد ثبت اختراع ۲۴ برابر طول به اسناد خبری است [۲]. مطالعه ساختار نحوی زبان ثبت اختراع نیز چالش بزرگی

1. World Intellectual Property Organization (WIPO)
2. Patent Application
3. Recall-Oriented Problem
4. Outlier Points

*نویسنده مسئول

است. نشان داده است نویسندگان پتنت تمایل دارند از عبارات چند کلمه‌ای برای معرفی اصطلاحات جدید استفاده کنند [۳]. چالش دیگر در جستجوی پتنت، مشکل عدم تطابق واژگان است، بدین معنی که یعنی عدم وجود کلمات مشترک بین دو سند مرتبط. مگدی و همکاران [۴] نشان دادند که ۱۲ درصد از اسناد پتنت در دادگان مربوط به CLEF-IP2009 در موضوعات مرتبط، هیچ کلمه مشترکی ندارند. موارد فوق، جستجوی ثبت اختراع را به یک فرایند پیچیده تبدیل می‌کند.

محققان روش‌های جستجو و بازیابی پتنت را دسته‌بندی کرده‌اند. لوپو و هانبری [۵] روش‌هایی را برای بازیابی پتنت خلاصه کردند که به روش‌های مبتنی بر متن (سبد کلمات، تحلیل معنایی پنهان، پردازش زبان طبیعی)، روش‌های مبتنی بر تغییر پرس و جو، روش‌شناسی مبتنی بر فراداده، و روش‌شناسی مبتنی بر طراحی تقسیم می‌شوند. در تحقیق انجام شده در [۶] روش‌های بازیابی پتنت به روش‌های مبتنی بر IPC، روش‌های مبتنی بر ویژگی‌های پتنت و روش‌های مبتنی بر ساخت پرس و جو تقسیم کرده‌اند. اخیراً، شلبی و همکاران در [۷] بازیابی پتنت را به دسته‌های زیر تقسیم کرده‌اند: روش‌های مبتنی بر کلمه کلیدی، روش‌های بازخورد شبه مرتبط[1]، روش‌های مبتنی بر معنا، روش‌های مبتنی بر فراداده و روش‌های تعاملی.

۱-۱- ساختار تقاضانامه ثبت اختراع

برای دریافت گواهی ثبت اختراع لازم است تقاضانامه ثبت پتنت به اداره ثبت اختراع در کشور موردنظر ارائه گردد. از آنجا که اجزای این تقاضانامه می‌تواند پرس و جوی موردنظر را شکل دهد، لذا در اینجا به اختصار ساختار اطلاعاتی تقاضانامه ثبت پتنت مورد بررسی قرار می‌گیرد.

بخش شرح اختراع[2]: بخش اصلی تقاضانامه را شرح اختراع تشکیل می‌دهد که در آن باید شرایط اصلی ثبت پتنت اثبات گردد. این بخش شامل عنوان اختراع، زمینه فنی اختراع[3]، دانش قبلی مربوط به اختراع[4] (آنچه از قبل در حوزه دانش بشر در مورد اختراع وجود دارد) و همچنین ارزیابی دانش فنی موجود[5] می‌باشد.

بخش افشای اختراع[6]: شامل مشخصات کامل اختراع موردنظر از لحاظ فنی و توصیف کامل و واضح راه‌حل ارائه‌شده، شرح تصاویر، شرح کاربرد صنعتی اختراع و قابلیت تولید اختراع در یک خط تولید و یا یک کارخانه، تأثیرات سودمند اختراع و نحوه پیاده‌سازی اختراع است.

بخش ادعانامه[7]: در این بخش حدود و ثغور قانونی حمایت درخواستی تعیین و موضوع اختراع و ویژگی‌های فنی و اساسی آن تشریح می‌گردد.

بخش چکیده[8]: چکیده شامل خلاصه فشرده‌ای از اختراع است و مشکل موجود و راه‌حل ارائه‌شده توسط اختراع را به صورت خلاصه تشریح می‌نماید. چکیده پس از ثبت پتنت به جستجوی پتنت موردنظر در بانک‌های اطلاعاتی کمک می‌نماید. این بخش صرفاً جهت استفاده فنی بوده و فاقد کاربرد حقوقی است.

بخش اشکال[9]: این بخش شامل شکل‌ها جداول و نمودارهای مورد نیاز جهت ثبت اختراع است.

۱-۲- چالش‌های بازیابی پتنت

در جستجو و بازیابی پتنت باید شرایط خاصی را موردنظر قرار داد که آن را از دیگر روش‌های جستجو مانند جستجو در وب متمایز می‌سازد. از جمله این شرایط خاص می‌توان به ویژگی‌های اسناد هدف اشاره نمود. برخی از خصوصیات اسناد پتنت را می‌توان به شرح ذیل برشمرد:

- نقطه آغازین جستجوی پتنت و یکی از چالش‌های اساسی، تبدیل سند تقاضانامه ثبت پتنت به یک پرس و جوی مؤثر و کارآمد است. به عبارت دیگر سند پتنت باید به‌عنوان نیاز اطلاعاتی مورد استفاده قرار گیرد. در این راستا تحقیقات مختلفی براساس فرکانس رخداد واژه‌های[10] موجود در سند تقاضانامه موردنظر انجام پذیرفته است.
- مجموعه واژگان مورد استفاده در یک سند تقاضانامه ثبت پتنت عموماً بسیار خاص، انحصاری و فنی هستند و معمولاً در مکالمات و نوشتار روزمره مورد استفاده قرار نمی‌گیرند. نویسندگان درخواست ثبت از عبارات کاملاً خاصی استفاده می‌کنند تا حوزه ادعای خود را بسیار محدود نمایند. این موجب می‌شود که به دلیل محدودیت واژگان مورد استفاده، گاه تطابقی بین سند و پرس و جو ایجاد نشود ولی مفهوم مندرج در پتنت بسیار شبیه یا حتی برابر با موضوع پرس و جو باشد. به عبارت دیگر، عموماً اختراعات جدید کلمات جدیدی را به حوزه فناوری وارد می‌کنند و استفاده گسترده از اختصارات و کلمات تکنیکی جدید چالشی اساسی برای سیستم‌های بازیابی اطلاعات است.
- سبک نگارش[11] در بخش‌های مختلف یک پتنت می‌تواند متفاوت باشد. به‌طور مثال بخش «زمینه فنی اختراع» با یک نوع شیوه نگارش و بخش حقوقی با سبک نگارش کاملاً متفاوتی نوشته می‌شود.
- یکی از خصوصیات اسناد پتنت آن است که برخلاف یک گزارش فنی که بسیار شفاف نوشته می‌شود، تأکید نویسنده پتنت آن است که چگونه حوزه پوشش پتنت خود را گسترش دهد و در عین حال خواننده متن نتواند به راحتی تکنیک‌های موردنظر در پتنت را

1. Pseudo Relevance Feedback Methods
2. Patent Description Section
3. Technical Field
4. Prior Art
5. Assessment of Prior Art
6. Disclosure Section
7. Claim Section
8. Abstract Section
9. Drawing Section
10. Term Frequency
11. Writing Style

درک نماید. به عبارت دیگر ایجاد ابهام در نگارش یکی از مهارت‌های متخصصین ثبت پتنت است. به‌عنوان مثال ممکن است به جای واژه «فنر» از عبارت «مفتول فلزی دوار» استفاده شود.

- یک سیستم بازیابی پتنت باید توانمندی اجرای «پرس و جوهای خاص» را داشته باشد. به‌عنوان مثال در جستجوی توصیفگرهای عددی «اتومبیل با ۵ چرخ»، بخش متمایزکننده این عبارت عدد ۵ است. حال آنکه اغلب سیستم‌های بازیابی، اعداد و حتی شکل حرفی اعداد را به صورت ایست واژه[1] در نظر می‌گیرند. همچنین در جستجوی جستجوی پرس و جوهای منفی مانند «شوینده بدون سفیدکننده» اغلب سیستم‌های بازیابی، تمامی اسناد مرتبط با سفیدکننده را باز می‌گردانند.

جستجو به دنبال پتنت‌های مشابه را اصطلاحاً جستجوی سنجش عدم اعتبار[2] می‌نامند. به عبارت دیگر هدف جستجو، یافتن پتنت‌هایی است که پرس و جوی موردنظر را غیرمعتبر می‌نمایند. متخصصین آزمونگر پتنت[3] بطور معمول یکصد تا دویست پتنت بازیابی‌شده توسط موتور جستجو را به دقت مورد ارزیابی قرار می‌دهند. حال آنکه روش‌های معمول بازیابی عمدتاً بر روی دقت بازیابی صفحه اول تأکید زیادی دارند. از این‌رو روش‌های سنتی به شیوه ساده قابل اعمال بر روی سامانه‌های جستجوی پتنت نیستند. نکته اساسی آن است که در بازیابی اطلاعات پتنت عمدتاً هدف آن است تا مسئولیت جستجو از سوی متخصصین آزمونگر پتنت به سیستم منتقل شده و این امر حتی‌الامکان به صورت خودکار انجام پذیرد.

۱-۳- ساختار مقاله

این مقاله به مروری بر سیستم‌ها و روش‌های بازیابی پتنت با استفاده از شیوه‌های نوین بازیابی هوشمند اطلاعات می‌پردازد. در بخش دوم ساختار و ویژگی‌های پایگاه‌های ثبت پتنت که جستجو باید در آنها صورت گیرد تشریح می‌گردد. بخش سوم به بازیابی تک زبانه اسناد پتنت و الگوریتم‌ها و شیوه‌های رایج در این خصوص می‌پردازد. در بخش چهارم شیوه‌های بازیابی بین زبانی اسناد پتنت تشریح می‌گردد. در بخش پنجم شیوه‌های ارزیابی سامانه‌های جستجوی پتنت شامل پیکره‌های آزمون و معیارهای ارزیابی دقت بازیابی مورد بررسی قرار می‌گیرند. در نهایت بخش ششم به بحث و نتیجه‌گیری در مورد شیوه‌های مختلف بازیابی هوشمند پتنت اختصاص دارد.

۲- ویژگی و ساختار پایگاه‌های ثبت پتنت

به منظور ایجاد یک جستجوی مؤثر و کارآمد، می‌بایست ساختار واحدهای اطلاعاتی در پایگاه ثبت پتنت مورد بررسی قرار گیرد. ساختار اطلاعاتی یک پتنت در اداره ثبت پتنت آمریکا[4] در جدول ۱ آمده است.

همانگونه که اشاره شد، پتنت امکان آن را فراهم می‌سازد تا مخترع در ازای پرداخت وجه، اختراع خود را افشا نموده و از مزایای حفاظت اختراع خود برخوردار گردد. در عین حال مخترع گاه تلاش می‌کند تا برای اطمینان بیشتر و جلوگیری از کپی‌سازی، مفاهیم اختراع خود را در مستندات حجیم و غیرقابل فهم پنهان سازند. بنابراین درحالی‌که یک اختراع را می‌توان به راحتی در تنها چند سطر توصیف نمود، بسیار طبیعی است که مخترع آن را در چندین صفحه توصیف نماید و ابهام را در متون افزایش دهد. این بدان معنی است که مدل‌های استاندارد بازیابی اطلاعات همچون مدل فضای برداری نمی‌تواند رویکرد مناسبی برای جستجوی اطلاعات پتنت باشد. لذا ساختارهای فراداده دیگری نیز برای دسترسی بهتر به اطلاعات موجود در پایگاه داده‌ها توسعه داده شده‌اند. در اینجا به شرح یکی از این ساختارهای سلسله مراتبی موضوعی می‌پردازیم.

جدول ۱- ساختار اطلاعاتی پتنت در اداره ثبت پتنت آمریکا (USPTO)

Ttl	Title
Abst	Abstract
Bsum	Background summary
Drwd	Description of the figures
Detd	Detailed description
Clms	Claims
Pclm	Primary claim

براساس طبقه‌بندی موجود در بانک اطلاعاتی پتنت‌ها، هر سند پتنت در یک ساختار موضوعی سلسله مراتبی تحت عنوان IPC[5] قرار گرفته است. این ساختار بالغ بر ۷۰۰۰۰ زیر بخش دارد و توصیف موضوعی مناسبی را از اختراع ارائه می‌نماید. تخصیص هر سند پتنت به هر یک از شاخه‌های این طبقه‌بندی توسط عوامل انسانی صورت گرفته و از این‌رو بسیار معتبر است. این طبقه‌بندی نیز می‌تواند به‌طور مؤثر در بازیابی اطلاعات پتنت مورد استفاده قرار گیرد.

از آنجا که به دلیل حجم گسترده و رو به رشد درخواست پتنت، برچسب‌گذاری تمامی پتنت‌ها و قراردادن آنها در ساختار طبقه‌بندی IPC به صورت دستی امری دشوار و زمان‌بر است، لذا بکارگیری روش‌های نوین بازیابی اطلاعات متنی بسیار کارگشا خواهد بود.

بنابراین برخی سیستم‌ها با وقوف بر این مسأله، تأکید دارند تا از الگوی طبقه‌بندی موجود استفاده نکرده و اطلاعات ذاتی موجود در بخش‌های مختلف متنی شناسنامه پتنت را جهت بازیابی به‌کار گیرند [۸]. در عمل، ترکیبی از روش‌های بازیابی هوشمند متون و اطلاعات ساختاری در بازیابی پتنت به‌کار می‌رود.

1. Stop Word
2. Invalidity Search Run
3. Patent Examiners
4. US Patent Office (USPTO)
5. International Patent Classification

۳- بازیابی تک زبانه اسناد پتنت

بازیابی تک زبانه اسناد پتنت معمولاً با سه چالش اصلی همراه است. چالش اول نحوه ساخت عبارت یا عبارت‌های پرس و جو از تقاضانامه ثبت پتنت است. چالش دوم نحوه جستجو در اسناد پتنت می‌باشد و در نهایت مشکل سوم آن است که چگونه اطلاعات ساختاری و اَبَرداده[1] برای کوچک‌کردن محدوده جستجو مورد استفاده قرار گیرد.

در طراحی سیستم‌های بازیابی اطلاعات پتنت برخی سیستم‌های جستجو بر روی توسعه مدل بازیابی تأکید کرده‌اند. حال آنکه نحوه ساخت یک پرس و جوی[2] مناسب نیز بسیار حائز اهمیت است. در تولید یک پرس و جو، سؤالات مهمی مطرح می‌شود؛ از جمله آنکه شیوه قطعه‌بندی سند ورودی چگونه است، چند کلمه جستجو برای پرس و جو مناسب است. کلمات پرس و جو از کجا استخراج شود، چگونه وزن‌دهی شود، و چگونه از عبارات اسمی بهره‌گیری شود. همچنین تصحیح و بهبود پرس و جو نیز از اهمیت بالایی برخوردار است.

تحقیقات مختلفی بر روی استخراج واژگان پرس و جو در جستجوی پتنت انجام پذیرفته است. در تحقیق انجام‌شده در [۹] تمامی کلمات بخش «ادعانامه درخواست پتنت» به‌عنوان یک پرس و جوی طولانی مورد استفاده قرار گرفته است. دلیل ارائه‌شده از سوی محققین آن بوده که اولاً به دلیل پیچیدگی فرایند انتخاب، در صورت حذف واژگان امکان بروز خطا در سیستم وجود دارد و ثانیاً وزن‌دهی به کلمات با استفاده از روش TF-IDF[3] نوعی انتخاب واژه محسوب می‌گردد. بخش ادعانامه درخواست‌های پتنت با حذف ایست واژه‌ها به صورت یک پرس و جو مورد استفاده قرار گرفته است. به دلیل آنکه بسیاری از پتنت‌ها دارای بخش خلاصه نیستند، لذا تأکید این تحقیق بر روی بخش ادعانامه پتنت صورت گرفته است. دیدگاه محققین در این تحقیق آن است که بخش ادعانامه نسبت به دیگر بخش‌ها از غنای اطلاعاتی بیشتری برخوردار است. ارزیابی بازیابی براساس دو معیار MAP[4] و NDCG[5] انجام شده و نتایج مطلوبی در بر نداشته است. نمایه‌سازی و بازیابی در ابزار Lemur[6] انجام گرفته است.

در تحقیق انجام شده در [۱۰] تمامی سند پتنت بدون در نظر گرفتن اطلاعات ساختار به‌عنوان پرس و جو در نظر گرفته شده است. آزمایشات صورت گرفته در این تحقیق نشان داده است که فیلد background summary سودمندترین منبع اطلاعاتی برای ایجاد پرس و جو است. لازم به ذکر است این فیلد در پایگاه داده ثبت پتنت آمریکا معادل فیلد description در پایگاه ثبت پتنت اروپاست. از این‌رو نتیجه حاصل شده در این تحقیق هم‌راستا با پژوهش مندرج در [۱۱] است.

آزمایشات بدون حذف کلمات نویز و استفاده از ریشه‌یابی کلمات[7] صورت پذیرفته است. با بررسی تعداد کلمات انتخاب‌شده از هر فیلد بر روی دقت بازیابی از ۱۰ تا ۵۰ کلمه، نتایج بدست‌آمده نشان داده است که انتخاب ۱۰ کلمه از عنوان و ۲۰ کلمه از دیگر فیلدهای سند پتنت دقت بازیابی را بیشینه می‌نماید.

همانگونه که در جدول ۲ آمده است، استفاده از معیار وزن‌دهی TF و بکارگیری محتوای فیلد background summary بهترین نتیجه را ارائه نموده است.

در تحقیق انجام‌شده در [۱۲] به منظور افزایش اثربخشی بازیابی، از دسته‌بندی حوزه‌های فنی موجود در طبقه‌بندی بین‌المللی پتنت (IPC) برای تولید بهینه عبارت پرس و جو جهت محدودترکردن نتایج بازیابی و افزایش دقت آن بهره‌گیری می‌نمایند.

در پژوهشی دیگر با هدف تولید خودکار پرس‌وجو برای جستجو در پتنت، سه نوع از ویژگی‌های آماری و ساختاری در جستجو مورد بررسی قرار گرفته و ویژگی‌های مختلف با یک روش Learning To Rank با یکدیگر ترکیب شده‌اند [۱۳]. پایه این پژوهش استفاده از مدل پرس و جوی indri[8] است [۱۴]. موتور جستجوی indri هسته اصلی جستجو در پروژه Lemur می‌باشد. سه ویژگی مختلف در تبدیل تقاضانامه ثبت پتنت به پرس و جوی موردنظر مورد استفاده قرار گرفته‌اند. ویژگی‌های موردنظر در این تحقیق عبارتند از:

الف- ویژگی‌های مرتبط با امتیاز بازیابی[9]

برای اینکه یک پرس وجوی مؤثر از سند پتنت ساخته شود، به یک روش تبدیل نیازمندیم. به منظور طراحی روش تبدیل، فاکتورهای متعددی را می‌توان به‌کار گرفت. مجموعه این فاکتورها در جدول ۳ آمده است.

جدول ۲- تأثیر بکارگیری کلمات فیلدهای جستجو در دقت بازیابی [۱۰]

Field type	MAP		
	bool	TF	TF-IDF
Title	0.042	0.039	0.043
Description of the figures	0.044	0.048	0.047
Detailed description	0.055	0.057	0.066*
Primary claims	0.059	0.062	0.055
Claims	0.066	0.066	0.064
Abstract	0.066	0.070	0.074*
All	0.067	0.068	0.078*
Background summary	0.078	0.082	0.094*

1. Meta Data
2. Query Formulation
3. Term Frequency - Inverse Document Frequency
4. Mean Average Precision
5. Normalized Discounted Cumulative Gain
6. https://www.lemurproject.org
7. Stemming
8. https://www.lemurproject.org
9. Retrieval Score Features

جدول ۳- فاکتورهای مؤثردر ویژگی‌های مرتبط با بازیابی

فاکتور	توضیح	مقادیر مورد آزمایش در مقاله [۱۱]
Num	تعداد کلمات پرس و جو	• Between 10 to 100 words
Field	از کجا کلمات پرس و جو استخراج شود (شش فیلد)	• The title field (ttl), • Abstract field (abst), • Brief summary field (bsum), • Description of the figures(drwd), • Detailed text description field (detd) • Claim field (clms). • +Primary claim field (pclms),
weight	چه الگوی وزن‌دهی را بر روی کلمات اعمال کنیم	• Equal weight (bool), • Term frequency (TF) • Combination of TF and IDF (TF.IDF).
NP	استفاده از عبارات اسمی به‌عنوان مکمل	• Use noun-phrase (true) • Not to use noun-phrase (false).
All	استفاده از سند پتنت به‌عنوان متن پرس و جو	• All patent document as a query

ب - ویژگی‌های سطح پایین[1]

این ویژگی‌ها اشکال مختلف فرکانس رخداد واژه را در اسناد مورد استفاده قرار می‌دهد و شامل موارد ذیل هستند:

- TF
- Normalized (TF)
- Log (TF)
- IDF
- TF. IDF
- Log (TF). IDF
- Normalized (TF). IDF

این ویژگی‌های آماری عموماً در مدل پرس و جوی indri قابل بیان نیستند و از این جهت می‌توانند به بهبود کیفیت جستجو کمک نمایند.

ج - ویژگی‌های مرتبط با طبقه‌بندی[2]

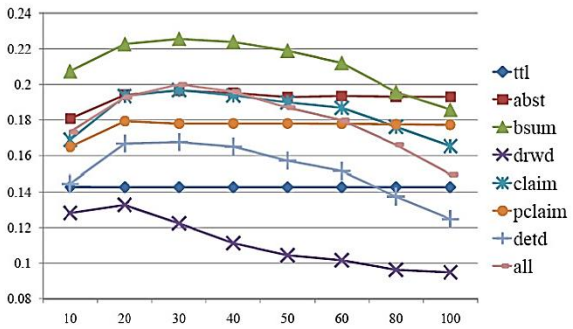
این ویژگی‌ها اطلاعات مرتبط با رده‌بندی موجود در بانک اطلاعاتی پتنت را مورد استفاده قرار می‌دهند. ویژگی مرتبط با طبقه‌بندی را می‌توان به صورت شباهت میان نشان‌های رده‌بندی در تقاضانامه ثبت پتنت و نشان‌های رده‌بندی در پایگاه داده‌های پتنت تعریف نمود. جدول ۴ ویژگی‌های مرتبط با طبقه‌بندی را در پژوهش انجام گرفته در [۱۳] به نمایش می‌گذارد.

جدول ۴- ویژگی‌های مرتبط با طبقه‌بندی

#	Category Type	Description
1	<OCL>	Primary Class Code
2	<XCL>	Secondary Class Code
3	<FSC/FSS>	Related Class Code
4	<ICL>	Class code of International Classification System

به منظور بهره‌گیری از این سه دسته ویژگی، ترکیب خطی آنها در تحقیق مورد استفاده قرار گرفته است. به منظور محاسبه وزن هریک از ویژگی‌ها در ترکیب کلی، از Adarank استفاده شده است که براساس محاسبه امتیاز بازیابی، به شکل خودکار وزن آن ویژگی خاص را محاسبه می‌نماید [۱۵].

نتایج نشان می‌دهد که ترکیب سه ویژگی مذکور کارایی جستجو را به میزان قابل ملاحظه‌ای بهبود می‌بخشد. همچنین بهترین ویژگی یکتا برای جستجو، ترکیب کلمات و عبارات اسمی حاصل از فیلد خلاصه[3] است. علاوه بر این شکل ۱ نشان می‌دهد که استفاده از فیلد خلاصه بسیار بهتر از فیلد ادعانامه عمل کرده و فیلد عنوان بدترین نتیجه را در بازیابی اطلاعات دارد.



شکل ۱- منحنی دقت بازیابی برحسب تعداد کلمات پرس و جو برای هریک از فاکتورهای جدول سه [۱۳]

بازیابی اطلاعات پتنت در حوزه شیمی و مهندسی شیمی نیز به‌طور ویژه مورد توجه قرار دارد. مجموعه پایگاه داده TREC-CHEM.2009 به منظور بررسی و مقایسه روش‌های بازیابی اطلاعات پتنت در حوزه شیمی توسعه یافته است. اصولاً پردازش و جستجوی اطلاعات حوزه شیمی فرایندی بسیار دشوار است. وجود اسامی متعدد و بسیار متفاوت، نام‌های مختلف تجاری و فرمول‌ها فرایند این جستجو را بسیار پیچیده می‌نماید. به‌عنوان مثال کلمه Aspirin بیست و پنج واژه مترادف و ۹۵ نام تجاری دارد. در تحقیقات انجام‌شده در این خصوص، استفاده از گسترش پرس و جو و همچنین استفاده از فهرست پتنت‌هایی که در بخش نقل قول[4] سند پتنت آمده است می‌تواند بسیار کارگشا باشد [۱۶،۱۷،۱۸].

در تحقیق انجام‌شده در [۱۱] به استخراج خودکار پرس و جو از مجموعه داده‌های سند پتنت پرداخته شده است. بدین ترتیب که توزیع واژگان در فیلدهای مختلف یک پتنت مورد ارزیابی قرار گرفته و با توزیع واژگان در کل پیکره بوسیله اعمال یک مدل زبانی[5] مقایسه می‌شود.

به منظور انجام مقایسه بین واژگان پرس و جو و مجموعه پیکره، از روش KL-Divergence بهره‌گیری شده است. این روش کلماتی را که در پرس و جو پرتکرار ظاهر شده و در مجموعه پیکره کم تکرار هستند تقویت می‌نماید. در این تحقیق از طبقه‌بندی IPC موجود در پتنت‌ها نیز استفاده شده است.

1. Low Level Features
2. Category Features

3. Summary Field
4. Citation Section
5. Language Model

در استخراج مدل پرس و جو ابتدا یک مدل براساس تخمین وزن‌دار log-likelihood ساخته می‌شود. شیوه ساخت مدل آن است که فرکانس نسبی کلمات در فیلدهای مختلف سند پتنت (title, description, abstract, claims) بدست می‌آید و سپس هموارسازی بر روی آن انجام می‌گیرد. همچنین به منظور افزایش دقت بازیابی یک برچسب از طبقه‌بندی‌های IPC به این مدل ضمیمه می‌شود تا دقت جستجو را بالا ببرد. نکته حائز اهمیت در عملکرد هموارسازی آن است که با این عمل دانش ضمنی موجود در طبقه‌بندی IPC به نحوی در فرایند جستجو لحاظ خواهد گردید. این امر به منزله گسترش مدل پرس و جو از طریق فراداده طبقه‌بندی IPC است. نتایج این تحقیق نشان می‌دهد که بخش توصیف پتنت[1] مهم‌ترین و اثربخش‌ترین بخش یک فایل پتنت برای استخراج واژگان جستجو است.

اصولاً در ساخت عبارت پرس و جو برای جستجو در مجموعه داده‌های پتنت، عبارات پرس و جو بسیار بزرگ هستند. از این‌رو علاوه بر شیوه‌های گسترش پرس‌وجو، از روش‌های کاهش اندازه پرس و جو نیز استفاده می‌شود [۱۹].

در سال‌های اخیر استفاده از شبکه‌های عصبی در بازیابی هوشمند اطلاعات و به تبع آن در بازیابی پتنت افزایش یافته است که یک زمینه جدید و در حال توسعه است [۲۰]. مدل‌های ترانسفورمری مانند BERT به نتایج چشمگیری در وظایف مختلف NLP دست یافته‌اند [۲۱]. در حال حاضر BERT توجه زیادی را در تحقیقات صنعت ثبت اختراع به خود جلب کرده است. بازیابی متراکم [۲۲] یک روش عصبی جدید برای جستجو است و با توجه به ویژگی‌های خاص حوزه ثبت اختراع، انتظار می‌رود مشکلاتی مانند عدم تطابق واژگان را حل کند و اثربخشی بازیابی را بهبود بخشد.

در تحقیق انجام‌شده توسط استاماتیس [۲۳]، کارایی روش‌های مبتنی بر شبکه‌های عصبی مانند BERT برای جستجوی اسناد اختراع مورد بررسی و ارزیابی قرار گرفته است. در این تحقیق، مدل BERT با ویژگی‌های ثبت اختراع تطبیق داده شده است تا عملکرد بازیابی افزایش یابد. این تحقیق از یک رویکرد بازیابی دو مرحله‌ای بهره‌گیری می‌نماید. در مرحله اول از الگوریتم BM25 استفاده شده و در گام بعد مدل BERT بر روی امتیاز BM25 عمل کرده و آن را اصلاح می‌کند. علت استفاده از رویکرد دو مرحله‌ای، کاهش حجم محاسبات است.

۴- بازیابی بین زبانی اسناد پتنت

بازیابی بین زبانی[2] به فرایندی اطلاق می‌شود که در آن زبان پرس‌وجو متفاوت از زبان اسناد باشد. در تحقیق انجام شده در [۸] ، سیستمی تحت عنوان bSmart در پایگاه پتنت اروپا (EPO) معرفی و ارائه شده است که به تحلیل زبان‌های انگلیسی، ژاپنی، آلمانی، فرانسه و همچنین جستجوی بین زبانی می‌پردازد. متداول‌ترین روش موجود برای یک سیستم بازیابی بین زبانی آن است که ابتدا عبارت پرس‌وجو را به زبان اسناد ترجمه کرده و سپس یک بازیابی تک زبانی انجام گیرد. استفاده از لغتنامه برای یافتن ترجمه‌های متفاوت از هر واژه موجب می‌شود تا ترجمه و همچنین گسترش پرس‌وجو با دقت بالایی انجام پذیرد. ولی مشکل آنجاست که بروزرسانی این لغت‌نامه از پیچیدگی خاصی برخوردار است. در حوزه ثبت پتنت این امر با کمک کاربران سیستم امکان‌پذیر خواهد بود. نکته قابل ذکر آنکه ترجمه می‌تواند قبل از گسترش پرس‌وجو و یا پس از آن صورت پذیرد.

یکی از مشکلات اسناد پتنت در بازیابی بین زبانی آن است که وکلای ثبت پتنت[3] عموماً در نوشتار خود از جملات و کلمات مبهم و کلی استفاده می‌کنند، زیرا استفاده از کلمات مشخص و بسیار خاص شفاف ممکن است حفاظت از پتنت را به خطر انداخته و حوزه حفاظت را محدود نماید. لذا وکلای ثبت پتنت معمولاً تمایل دارند تا از کلمات ناشفاف برای توصیف پتنت استفاده نمایند. به‌عنوان مثال به جای واژه «فنر» از عبارت «مفتول سیمی استوانه‌ای فشرده شونده در امتداد یک محور» استفاده می‌نمایند. در این حالات حتی فرایند گسترش پرس‌وجو نیز حجم زیادی از کلمات را ارائه نموده و فرایند جستجو و بازیابی بین زبانی را با پیچیدگی زیادی همراه می‌سازد [۸].

از مزایای پایگاه‌های اطلاعاتی ثبت پتنت آن است که به دلیل آنکه جملات با دقت فراوانی انتخاب شده است، لذا نوشتار از لحاظ گرامری کاملاً درست بوده و در آن جملات محاوره‌ای وجود ندارد. این امر به فرایند ترجمه و جستجو کمک شایانی خواهد نمود. افعال به صورت اول شخص و یا دوم شخص استفاده نمی‌شود و همچنین صرف فعل در زمان گذشته یا آینده صورت نمی‌گیرد. این موارد امکان آن را فراهم می‌سازد تا سیستم‌های تجزیه‌گر جملات[4] به خوبی و با صحت بالا عمل کنند.

هر زبان ویژگی‌های منحصر بفرد خود را در مواجهه با بازیابی اطلاعات دارد. زبان انگلیسی شامل لغات با نقش‌های متفاوت و مبهم است (مانند دو نقش اسم و فعل). زبان آلمانی در ارتباط با رایانه رفتار دوستانه‌تری دارد. زبان‌هایی مانند فرانسه و اسپانیایی به واسطه تعداد زیاد پسوندها و شیوه صرف افعال، مشکلات خاص خود را دارند. ساخت عبارات اسمی با استفاده از حروف اضافه نیز مشکلاتی را در ترجمه این زبان‌ها ایجاد می‌نماید.

با لحاظ نمودن تفاوت‌های بین زبانی، این تحقیق در سیستم bsmart مورد استفاده در [۸] از روش نمایه‌سازی عبارات (به ویژه عبارات اسمی[5]) به جای نمایه‌سازی کلمات استفاده نموده است. ریشه‌یابی بر روی کلمات صورت می‌گیرد و گزیده‌گویی در جملات تشخیص داده می‌شود. روش وزن‌دهی به عبارات جستجو به قرار زیر است.

$$W_i = pip_i \times IDF_i \quad \text{معادله (۱)}$$

1. Patent Description Section
2. Cross lingual Information Retrieval (CLIR)
3. Patent Attorneys
4. Parsers
5. Noun Phrase

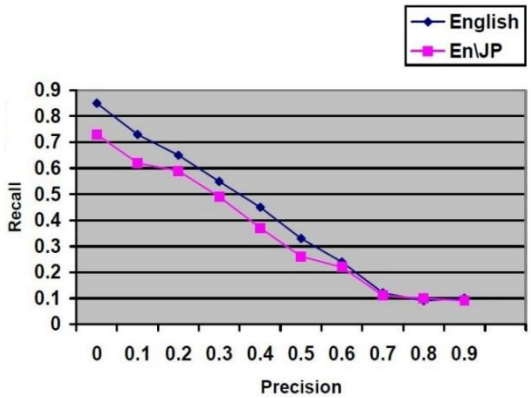
در این معادله W_i وزن واژه موردنظر، pip_i فاکتور جایگاه کلمه در عبارت[1] و IDF_i عکس فرکانس رخداد واژه در کل اسناد می‌باشد.

پارامتر pip در تعیین وزن کلمات از آن جهت استفاده می‌شود تا در عبارت موردنظر به برخی کلمات اهمیت بیشتری داده شود. به‌عنوان مثال در عبارت اسمی building block کلمه دوم و در عبارت اسمی block building نیز کلمه دوم بخش اصلی عبارت تلقی می‌گردد. در زبان انگلیسی عموما آخرین کلمه در یک عبارت اسمی به‌عنوان اسم اصلی[2] تلقی می‌شود و از این‌رو باید وزن بیشتری بدان تعلق گیرد. در زبان فرانسه و اسپانیایی بر خلاف انگلیسی، کلمه اول نام اصلی و دنباله عبارت، توصیف‌گر نام است. زبان ژاپنی ترکیبی از دو شکل انگلیسی و فرانسه است.

نکته حائز اهمیت آن است که در فرایند وزن‌دهی، پارامتر فرکانس رخداد واژه در سند مورد استفاده قرار نگرفته است. دلیل این امر آن است که همانگونه که پیشتر اشاره شد، معمولاً در یک پتنت، کلمات و مفاهیم اصلی اختراع در میان انبوهی از عبارات سند پتنت پنهان شده است.

نکته مهم دیگر آنکه ضریب IDF در رده‌های مختلف طبقه‌بندی، متفاوت در نظر گرفته شده است. لذا یک واژه در طبقه‌بندی‌های مختلف ممکن است IDFهای متفاوتی داشته باشد. این کار نتایج بازیابی را به طرز قابل ملاحظه‌ای بهبود می‌بخشد.

بردار حاصل از وزن‌دهی پس از هنجارسازی از طریق حاصل ضرب نقطه‌ای با عبارت پرس‌وجو مقایسه می‌گردد. آزمایشات بر روی پرس‌وجوهایی با متوسط طول ۲۳ کلمه انجام گرفته است. زبان پرس‌وجو انگلیسی و زبان بازیابی ژاپنی بوده است. نتیجه حاصل در شکل ۲ مشاهده می‌شود.



شکل ۲- منحنی PR در مقایسه بازیابی بین زبانی و بازیابی تک زبانی [۸]

منحنی PR[3] در شکل ۲ نشان می‌دهد که تفاوت معناداری میان بازیابی تک‌زبانی و بازیابی بین زبانی وجود ندارد. لذا می‌توان بدون تسلط به زبان ژاپنی، در اسناد پتنت‌هایی که به زبان ژاپنی نگارش شده‌اند جستجو نمود.

1. Position in Phrase
2. Head Noun
3. Precision-Recall Curve

۵- ارزیابی سامانه‌های جستجوی پتنت

۵-۱- پیکره دادگان

از سال ۲۰۰۷ سه کارزار ارزیابی برای مطالعه محک‌زنی روش‌های بازیابی پتنت در حوزه بازیابی هوشمند اطلاعات با نام NTCIR، CLEF-IP و TREC-CHEM آغاز گردید. در این فعالیت‌ها، مجموعه‌ای از دادگان مبنا برای بررسی صحت عملکرد سیستم‌های جستجوی پتنت توسعه داده شد.

مجموعه دادگان آموزشی مورد استفاده برای آزمون‌های جستجوی پتنت تحت عنوان CLEF-IP 2010 مشتمل بر ۳/۵ میلیون سند از اداره ثبت پتنت اروپا[4] است و حجمی معادل 75GByte دارد [۲۵،۲۴]. این پیکره چند زبانه بوده و حاوی اطلاعات پتنت در هریک از زبان‌های انگلیسی، فرانسه و آلمانی است. این پیکره اطلاعات طبقه‌بندی پتنت‌ها در فرمت IPC را نیز داراست. همچنین به تبع زمان ثبت پتنت، ویرایش‌های متفاوتی از پتنت‌ها در این پیکره یافت می‌شود. مجموعه دادگان آموزشی CLEF-IP علاوه بر پیکره فوق، یک مجموعه با ۲۰۰۰ عنوان موضوعی شامل ۶۰۰ هزار پتنت را نیز در بر دارد.

پیکره مورد اشاره به‌عنوان بستر اطلاعاتی بسیاری از تحقیقات از جمله در [۱۱] و [۹] مورد استفاده قرار گرفته است.

۵-۲- معیارهای ارزیابی دقت بازیابی

در تحقیقات صورت گرفته، طیف متنوعی از معیارها برای ارزیابی دقت بازیابی مورد استفاده قرار گرفته است. در تحقیق انجام‌شده در [۹] از دو معیار MAP و NDCG استفاده شده است. در پژوهش انجام‌گرفته در [۱۰] معیارهای MAP و Precision@10 برای ارزیابی به‌کار گرفته شده‌اند. (برای بسیاری از کاربردها به ویژه جستجوی وب، آنچه اهمیت دارد این است که چه تعداد نتایج مناسب در صفحه اول ظاهر می‌شود. لذا دقت اندازه‌گیری در چند سند اول از اهمیت بالایی برخوردار است. این به‌عنوان "دقت در k" یا Precision@k نامیده می‌شود. عموماً عدد k=10 در نظر گرفته می‌شود.)

همانگونه که قبلاً ذکر آن رفت بازیابی پتنت از جمله مسائل مرتبط با فراخوانی است. کیفیت سیستم‌هایی که به فراخوانی حساس هستند به مقدار زیادی مرتبط با شیوه مرتب‌سازی نتایج است. فراخوانی نرمالیزه‌شده[5] می‌توان معیار خوبی برای ارزیابی سامانه‌های حساس به فراخوانی باشد.

$$R_{norm} = 1 - \frac{\sum r_i - \sum i}{n(N-n)} \qquad \text{معادله (۲)}$$

4. European Patent Office (EPO)
5. Normalized Recall

در این معادله، r_i رتبه‌ای است که در آن i امین سند مرتبط بازیابی شده است. N تعداد کل اسناد در مجموعه مورد جستجو و n تعداد اسناد مرتبط است. ولی این معیار تنها برای مجموعه داده‌های با حجم کم مناسب می‌باشد. زیرا برای بدست‌آوردن آن باید کل اسناد مجموعه مرتب گردد.

معیار ارزیابی PRES[1] یکی از معیارهایی است که به صورت اختصاصی برای بازیابی اطلاعات مبتنی بر فراخوانی طراحی شده است و معیار جدیدی است که کیفیت سیستم‌های بازیابی پتنت را مورد ارزیابی قرار می‌دهد [۲۰،۲۶]. معیار PRES با اعمال تغییراتی بر روی معیار فراخوانی نرمالیزه‌شده بدست می‌آید. این معیار در حقیقت معیار فراخوانی[2] را با کیفیت رتبه‌بندی نتایج بازیابی ترکیب می‌نماید. به عبارت دیگر این معیار به ما این امکان را می‌دهد که بتوانیم کیفیت سیستم‌هایی که دارای معیار فراخوانی برابر یا نزدیک به هم هستند از یکدیگر تفکیک نماییم.

معادله (۳)
$$PRES = 1 - \frac{\frac{\sum r_i}{n} - \frac{n+1}{2}}{N_{\max}}$$

که در معادله فوق پارامتر $\sum r_i$ از عبارت زیر بدست می‌آید:

معادله (۴)
$$\sum r_i = \sum_{i=1}^{nR} r_i + nR(N_{\max} + n) - \frac{nR(nR-1)}{2}$$

در این معیار، کاربر حداکثر تعداد رکورد بازیابی مورد انتظار را تعیین می‌نماید. این معیار توانمندی سیستم را در بازیابی تمامی اسناد مرتبط می‌سنجد. بر خلاف MAP و Recall، این معیار به تلاشی که کاربران برای یافتن اسناد مرتبط به‌کار می‌برند وابسته است. این امر توسط یک پارامتر قابل تنظیم Nmax میسر می‌گردد که مقدار آن توسط کاربران تعیین می‌شود. این پارامتر حداکثر تعداد اسنادی است که کاربر مایل است در لیست مرتب‌شده اسناد بررسی نماید.

معیار PRES اثربخشی اسناد مرتب‌شده را برحسب بهترین و بدترین حالات رتبه‌بندی[3] می‌سنجد. بهترین حالت رتبه‌بندی آن است که تمامی اسناد مرتبط در بالای لیست قرار گیرند و بدترین حالت آن است که تمامی اسناد مرتبط درست بعد از Nmax (تعداد ماکزیمم اسنادی که باید بررسی شوند) واقع شوند. در حقیقت Nmax تعریف جدیدی را برای کیفیت مرتب‌سازی اسناد مرتبط ارائه می‌نماید.

اشکال دیگری نیز از معیار فوق‌الذکر نیز در تحقیقات بکار می‌رود. به‌عنوان مثال در تحقیق انجام‌شده در [۲۷] معیار PRES@100 مورد استفاده قرار گرفته است.

1. Patent Retrieval Evaluation Score (PRES)
2. Recall
3. Ranking

۶- نتیجه‌گیری

با افزایش حجم داده‌های ثبت اختراع در فضای وب و استفاده روزافزون از آن، بازیابی مؤثر اطلاعات در اسناد ثبت اختراع برای انجام فعالیت‌های نوآورانه امری ضروری است. با پیشرفت‌های اخیر فناوری، تجزیه و تحلیل پتنت نقش فزاینده‌ای در تعریف راهبردهای کسب‌وکارها دانش پایه ایفا می‌کند. این مقاله به بررسی و مرور ادبیات و تکنیک‌های مبتنی بر متن کاوی برای تجزیه و تحلیل پتنت و طبقه‌بندی آن ارائه می‌کند. بررسی ادبیات و پیشینه موضوع بر این واقعیت صحه می‌گذارد که حق ثبت اختراع یک سند خاص بوده و بازیابی آن یک امر چالش برانگیز است. بازیابی هوشمند اطلاعات با توجه به ویژگی‌های خاص پایگاه‌های داده پتنت و همچنین ابعاد حقوقی پیچیده آن، می‌تواند به صورت مؤثری به متخصصین آزمون‌گر پتنت جهت انجام جستجو یاری نماید.

مدل‌ها، الگوریتم‌ها و تکنیک‌های مختلف بازیابی اطلاعات توسط محققان پیشنهاد شده‌اند، اما هیچ تکنیک واحدی برای بازیابی پتنت مؤثر نیست و می‌بایست ترکیبی از آنها بکار گرفته شود. مطالعات بر روی فرمول‌های پرس و جوی ثبت اختراع با استفاده از تکنیک‌های بسط پرس‌وجو به ندرت افزایش مؤثری را در بازیابی نشان داده است. استفاده از IPC در پس پردازش ممکن است نتایج بهتری را برای رتبه‌بندی و فیلترکردن در صورت ترکیب با روش‌های دیگر استفاده از متن پتنت به همراه داشته باشد. با توجه به تکنیک‌ها و چارچوب‌های مختلف موجود و محدودیت‌های آنها، دامنه زیادی در زمینه تکنیک‌های بازیابی پتنت وجود دارد که فضای مناسبی را برای تحقیقات بیشتر در این حوزه ایجاد می‌کند.

درخصوص انجام تحقیقات آتی در زمینه بازیابی هوشمند پتنت، موارد مختلفی می‌تواند به‌عنوان حوزه‌های پیشنهادی تحقیق در حوزه بازیابی اطلاعات اسناد پتنت موردنظر قرار گیرد. از جمله این موارد می‌توان به استفاده از ساختار نقل قول‌ها[4]، استفاده از عبارات اسمی و یا بکارگیری اطلاعات یک زیربخش از طبقه‌بندی IPC اشاره نمود. به‌طورکلی ترکیب اطلاعات غیر متنی با روش‌های مبتنی بر بازیابی هوشمند اطلاعات می‌تواند در کوچک‌کردن محدوده جستجو و افزایش کیفیت بازیابی پتنت بسیار مؤثر باشد. نکته دیگر حائز اهمیت در بازیابی پتنت آن است که علاوه بر جستجو و بازیابی اسناد پتنت، جستجوی قسمت‌های کوتاه متنی[5] متنی[5] در داخل یک سند پتنت نیز می‌تواند به کاربر در یافتن بخش‌های اساسی در سند اختراع یاری رساند.

4. Citations
5. Passages

۷- مراجع

1- Hall, B. H. Patents and patent policy. Oxford Review of Economic Policy, 23(4), 568-587, 2007.
2- Iwayama, Makoto, Atsushi Fujii, Noriko Kando, and Yuzo Marukawa. "An empirical study on retrieval models for different document genres: patents and newspaper articles." In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 251-258. 2003.
3- Verberne, S., D'hondt, E., Oostdijk, N., Koster, C.H.: Quantifying the challenges in parsing patent claims. In: 1st International Workshop on Advances in Patent Information Retrieval (2010).
4- Magdy, W., Leveling, J., Jones, G.J.F.: Exploring structured documents and query formulation techniques for patent retrieval. In: Peters, C., et al. (eds.) Multilingual Information Access Evaluation I. Text Retrieval Experiments. CLEF 2009. Lecture Notes in Computer Science, vol. 6241. Springer, Berlin, Heidelberg.
5- Lupu, M., Hanbury, A.: Patent retrieval. Found. Trends Inf. Retreival 7(1), 1–97 (2013).
6- Khode, Alok, and Sagar Jambhorkar. "A literature review on patent information retrieval techniques." Indian Journal of Science and Technology 10, no. 36 (2017): 1-13.
7- Shalaby, W., Zadrozny, W.: Patent retrieval: a literature review. Knowl. Inf. Syst. 61(2), 631–660 (2019).
8- L. Sarasúa, "Cross Lingual issues in patent retrieval" SIGIR'00, pp. 1-4, 2000.
9- S.Verberne and E.D'hondt, "Prior art retrieval using the claims section as a bag of words", CLEF'09, pp. 1-3, 2009.
10- X. Xue and W. B. Croft, "Transforming Patents into Prior-Art Queries", SIGIR'09, pp. 1-2, 2009.
11- P. Mahdabi, M. Keikha, S. Gerani, M. Landoni, and Crestani, "Building Queries for Prior-art Search", IRFC'11, pp. 1-14, 2011.
12- Khode, Alok, and Sagar Jambhorkar. "Effect of technical domains and patent structure on patent information retrieval." International Journal of Engineering and Advanced Technology 9.1 (2019): 6067-6074.
13- Xiaobing Xue, and W. Bruce Croft, "Automatic Query Generation for Patent Search" CIKM'09, November 2–6, 2009, Hong Kong, China.
14- Strohman, Trevor, Donald Metzler, Howard Turtle, and W. Bruce Croft. "Indri: A language model-based search engine for complex queries." In Proceedings of the International Conference on Intelligent Analysis, vol. 2, no. 6, pp. 2-6. 2005.
15- Xu, Jun, and Hang Li. "Adarank: a boosting algorithm for information retrieval." In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 391-398. 2007.
16- H. Gurulingappa, B. Muller, M. Hofmann-Apitius, R. Klinger, H. T. Mevissen, C. M. Friedrich, J. Fluck "Prior Art Search in Chemistry Patents Based On Semantic Concepts and Co-Citation Analysis", The Nineteenth Text REtrieval Conference (TREC 2010) Proceedings.
17- J. Gobeill, A. Gaudinat, P. Ruch, E. Pasche, D. Teodoro, D. Vishnyakova, " BiTeM site Report for TREC Chemistry 2010: Impact of Citations Feeback for Patent Prior Art Search and Chemical Compounds Expansion for Ad Hoc Retrieval", The Nineteenth Text REtrieval Conference (TREC 2010) Proceedings.
18- H. Gurulingappa, B. Müller, R. Klinger, H.-T. Mevissen, M. Hofmann-Apitius, J. Fluck and C.M. Friedrich,"Patent Retrieval in Chemistry Based on Semantically Tagged Named Entities", The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings.
19- Bouadjenek, Mohamed Reda, Scott Sanner, and Gabriela Ferraro. "A Study of Query Reformulation for Patent Prior Art Search with Partial Patent Applications." In Proceedings of the 15th International Conference on Artificial Intelligence and Law, pp. 23-32. ACM, 2015.
20- Piroi, Florina, Mihai Lupu, Allan Hanbury, Alan P. Sexton, Walid Magdy, and Igor V. Filippov. "CLEF-IP 2010: Retrieval Experiments in the Intellectual Property Domain." In CLEF (notebook papers/labs/workshops). 2010.
21- Lupu, Mihai, and Allan Hanbury. "Patent Retrieval." Foundations and Trends in Information Retrieval 7, no. 1 (2013): 1-97.
22- W. Magdy, and Jones, G. "A new metric for patent retrieval evaluation". First International Workshop on Advances in Patent Information Retrieval (AsPIRe'10) at 32nd European Conference on Information Retrieval (ECIR 2010), 28 March 2010, Milton Keynes, U.K.
23- Magdy, Walid, and Gareth JF Jones. "PRES: a score metric for evaluating recall-oriented information retrieval applications." In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 611-618. ACM, 2010.
24- W.Magdy, P. Lopez, and G. J. F. Jones, "Simple vs. Sophisticated Approaches for Patent Prior-Art Search", ECIR'11, pp. 1-4, 2011.
25- Stamatis, Vasileios. "End to End Neural Retrieval for Patent Prior Art Search." In European Conference on Information Retrieval, pp. 537-544. Cham: Springer International Publishing, 2022.
26- Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. In: Empirical Methods in Natural Language Processing (EMNLP) (2020).
27- Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in arXiv: 1810.04805v2 (2019).