دانشگاه اصفهان

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر

پایان نامه‌ی کارشناسی ارشد رشته‌ی مهندسی کامپیوتر گرایش نرم افزار

گسترش معنایی پرس‌وجو

استاد راهنما:

دکتر محمد علی نعمت بخش

استاد مشاور:

دکتر ناصر نعمت بخش

پژوهشگر:

مژگان شبان زاده حبیب آبادی

مهرماه ۱۳۸۹

دانشگاه اصفهان

دانشکده فنی و مهندسی

گروه مهندسی کامپیوتر

پایان نامه ی کارشناسی ارشد رشته ی مهندسی کامپیوتر گرایش نرم افزار

خانم مژگان شبان زاده حبیب آبادی تحت عنوان

گسترش معنایی پرس‌وجو

در تاریخ ۸۹/۷/۴ توسط هیأت داوران زیر بررسی و با درجه عالی به تصویب نهایی رسید.

۱- استاد راهنمای پایان نامه دکتر محمد علی نعمت بخش با مرتبه ی علمی دانشیار امضا

۲- استاد مشاور پایان نامه دکتر ناصر نعمت بخش با مرتبه ی علمی استادیار امضا

۳- استاد داور داخل گروه دکتر مخمدرضا خیامباشی با مرتبه ی علمی استادیار امضا

۴- استاد داور خارج از گروه دکتر سید رسول موسوی با مرتبه ی علمی استادیار امضا

امضای مدیر گروه

دانشکده فنی مهندسی

سپاسگزاری

بر خود لازم می‌دانم تا از استاد راهنمای ارجمندم جناب آقای دکتر محمدعلی نعمت بخش و همچنین استاد مشاور گرانقدرم جناب آقای دکتر ناصر نعمت بخش کمال تشکر و قدردانی را داشته باشم و از خداوند توفیق ایشان را طلب می‌کنم.

از خانواده مهربانم که همواره مشوق و حامی اینجانب در طول زندگی بوده‌اند، سپاسگزاری کرده، سلامتی و موفقیت آن‌ها را از پروردگار مهربان مسئلت دارم. همچنین از تمامی دوستانی که در این راه یاریگر اینجانب بوده‌اند مراتب تشکر و قدر دانی را دارم.

تقدیم به

پدر و مادرم که موفقیت من آروزی آن‌ها و سربلندی آن‌ها

آرزوی من است.

چکیده

بازیابی اطلاعات یکی از اصلی‌ترین نیازهای کاربران است؛ روزانه کاربران زیادی به جستجو در وب و دیگر منابع به منظور پاسخ گویی به نیاز اطلاعاتی خود می‌پردازند. مسائل موجود در زبان طبیعی از جمله عدم تطابق لغوی، کلمات چند معنایی، کوتاه و مبهم بودن پرس‌وجو و دانش ناقص کاربران از موضوع مورد نظر در بازیابی اطلاعات منجر به بازیابی نتایج نامرتبط و کاهش رضایت کاربران از نتایج بازیابی شده می‌شود. گسترش پرس‌وجو با بررسی پرس‌وجوهای کاربران و افزودن خودکار کلمات مناسب و با ارزش به آن‌ها کمک می‌کند تا اسناد مرتبط با نیاز و منظور کاربر جستجو و بازیابی گردد. اگر گسترش پرس‌وجو به صورت هوشمندانه‌ای انجام نشود، با انحراف پرس‌وجو و فاصله گرفتن آن از منظور کاربر منجر به بازیابی نتایج نامرتبط‌تری نسبت به نتایج پرس‌وجوی اولیه خواهد شد. مساله‌ی دیگری که در گسترش پرس‌وجو باید به آن توجه کرد این است که در مورد پرس‌وجوهای مشتمل بر بیش از یک کلمه، انتخاب واژگان گسترشی که تنها با یکی از این کلمات مرتبط باشند باعث رخداد مشکل خروج از تعادل پرس‌وجو خواهد شد.

در این رساله، روش جدیدی برای گسترش معنایی پرس‌وجو به منظور تطبیق دادن پرس‌وجو با منظور کاربر ارائه شده‌است. روش پیشنهادی با استفاده از یک الگوریتم رفع ابهام مبتنی بر هستی‌شناسی به رفع ابهام از کلمات پرس‌وجو می‌پردازد. سپس، به منظور در نظر گرفتن روابط بین لغات در پرس‌وجوهای چند کلمه‌ای و اجتناب از مشکل خروج از تعادل پرس‌وجو به گروه‌بندی کلمات آن بر مبنای تشابه معنایی بین آن‌ها می‌پردازد. در ادامه با استفاده از روابط موجود در شبکه واژگان، یک شبکه‌ی معنایی از واژگان هر گروه ایجاد شده از کلمات پرس‌وجو و لغات مرتبط با آن‌ها از نظر معنایی می‌سازد. این روش بر طبق روابط و سلسله مراتب شبکه‌ی ساخته شده، مهمترین کلمات برای گسترش پرس‌وجو را مشخص می‌کند. از بین کلمات انتخاب شده، کلماتی که باعث ایجاد ابهام و نویز در پرس‌وجو نشوند، به عنوان کلمات گسترش انتخاب می‌شوند و وزن مناسبی برای آن‌ها محاسبه می‌شود. به این ترتیب پرس‌وجوی گسترش یافته ساخته می‌شود و عملیات جستجو با این پرس‌وجوی جدید انجام می‌شود. این روش با در نظر گرفتن معیارهای فراخوانی و دقت بر روی مجموعه داده‌ی TIME ارزیابی شده است. نتایج ارزیابی نشان دهنده‌ی افزایش نرخ فراخوانی و دقت بازیابی می‌باشد.

کلمات کلیدی: بازیابی اطلاعات، موتور جستجو، گسترش پرس‌وجو، روابط معنایی.

فهرست مطالب

عنوان **صفحه**

عنوان | صفحه

عنوان | صفحه

فهرست شکل‌ها

عنوان **صفحه**

عنوان **صفحه**

فهرست جدول‌ها

عنوان **صفحه**

فصل اول

مقدمه

۱-۱ مقدمه

فرآیند تشخیص ارتباط بین اسناد و پرس‌وجوی کاربر و بازگرداندن اسنادی که احتمالا نیاز کاربر را تامین می‌کنند، بازیابی اطلاعات نامیده می‌شود. بارزترین کاربرد سیستم‌های بازیابی اطلاعات موتورهای جستجو و کتابخانه‌های دیجیتالی می‌باشد. اگرچه با وجود موتورهای جستجو و کتابخانه‌های دیجیتالی دسترسی به اطلاعات آسان‌تر شده است، اما از آن جایی که حجم اطلاعات وب و کاربران آن روز به روز در حال افزایش می‌باشند، دسترسی به اطلاعات مورد نیاز به چالشی برای کاربران تبدیل شده است. علاوه براین، موتورهای جستجوی کلیدواژه‌ای در پاسخ به نیاز کاربران ضعیف عمل می‌کنند و بسیاری از کاربران نمی‌دانند چگونه پرس‌وجوی مناسب ایجاد کنند. حتی وقتی کاربران باتجربه در حوزه‌ی ناشناخته‌ای به جستجو می‌پردازند، قادر به ارائه‌ی پرس‌وجوی مناسب و بیان دقیق نیاز اطلاعاتی خویش نمی‌باشند [۱].

گسترش پرس‌وجو با افزودن کلمات به پرس‌وجو نتایج بازیابی اطلاعات را بهبود می‌بخشد. در این پایان نامه با ارائه‌ی روش جدیدی برای گسترش پرس‌وجو سعی می‌شود، بازیابی اطلاعات بهبود بخشیده شود.

در این فصل به معرفی سیستم‌های بازیابی اطلاعات، مشکلاتی که سیستم‌های بازیابی اطلاعات کلید واژه‌ای با آن مواجه می‌باشند و شرح مساله‌ی پژوهشی در این پایان‌نامه پرداخته می‌شود. سپس اهمیت و کاربرد موضوع بررسی

می‌شود و همچنین اهداف پایان‌نامه تبیین خواهد شد.

۱-۲ شرح و بیان مساله‌ی پژوهشی

بازیابی اطلاعات با مسائلی از قبیل نمایش، ذخیره، سازمان دهی و دسترسی اطلاعات مواجه می‌باشد [۲]. نمایش و سازمان دهی اطلاعات باید به گونه‌ای باشد که دسترسی آسان به اطلاعات دلخواه را برای کاربران فراهم کند. بیان دقیق و توصیف صفات اختصاصی نیاز اطلاعاتی[1] کاربران مساله‌ی ساده‌ای نیست. برای مثال نیاز اطلاعاتی زیر در وب را در نظر بگیرید.

> چرا تابع آزادسازی حافظه[2] در زبان برنامه‌نویسی C++ در محیط ویژوال بر روی سیستم‌عامل لینوکس با خطای دسترسی مواجه می‌شود؟

بدیهی است که این توضیح کامل از نیاز اطلاعاتی را نمی‌توان مستقیما برای درخواست اطلاعات با استفاده از واسط‌های کاربری موتورهای جستجوی امروزی به کار برد. بنابراین کاربر باید در ابتدا نیاز اطلاعاتی خود را به پرس‌وجویی قابل پردازش توسط موتورهای جستجو تبدیل کند.

معمولا خلاصه‌ای از نیاز اطلاعاتی کاربر توسط پرس‌وجویی مشتمل بر مجموعه‌ای از کلمات کلیدی بیان می‌شود. نیاز اطلاعاتی مطرح شده در مثال بالا برای یک موتور جستجوی انگلیسی زبان ممکن است به صورت یکی از پرس‌وجوهای زیرمطرح شود:

- free error access violation
- error at runtime and break
- "free function" access violation
- "free function" _CrtIsValidHeapPointer(pUserData)
- linux free error
- free function runtime error

با در اختیار داشتن پرس‌وجوی کاربر وظیفه‌ی اصلی سیستم‌های بازیابی اطلاعات، بازیابی اطلاعاتی است که مرتبط با نیاز اطلاعاتی کاربر باشند. ذکر این نکته ضروری است که سیستم‌های بازیابی اطلاعات متفاوت از سیستم‌های بازیابی داده می‌باشند.

---

[1] Information Need
[2] Free Function

۱-۲-۱ مقایسه‌ی سیستم‌های بازیابی اطلاعات با سیستم‌های بازیابی داده

در جدول ۱-۱ سیستم‌های بازیابی اطلاعات و سیستم‌های بازیابی داده با هم مقایسه شده‌اند. در سیستم‌های بازیابی داده مانند پایگاه داده‌های رابطه‌ای[1] داده‌ها ساختار یافته‌اند و از پرس‌وجوهای ساختار یافته برای بازیابی داده استفاده می‌شود. در صورت بازیابی پاسخ غلط در این سیستم‌ها، کل سیستم دچار خطا می‌باشد [۲]. در حالی که در سیستم‌های بازیابی اطلاعات، داده‌ها ساختار نیافته می‌باشند به این معنا که داده‌ها متنی می‌باشند و ساختار واضح و آسانی برای کامپیوترها ندارند. علاوه بر این در این سیستم‌ها پرس‌وجوها نیز غیر ساختار یافته و اغلب به زبان طبیعی می‌باشند. در نتیجه ممکن است نتایجی بازیابی شود که با پرس‌وجوی کاربر ارتباطی ندارند و یا ارتباط کمی دارند. چون پرس‌وجوها و داده‌ها به زبان طبیعی می‌باشند، زبان طبیعی ساختار نیافته است و ممکن است ابهام داشته باشند [۲].

جدول ۱- ۱ مقایسه‌ی سیستم‌های بازیابی اطلاعات با سیستم‌های بازیابی داده

سیستم‌های بازیابی اطلاعات	سیستم‌های بازیابی داده
داده‌های ساختار نیافته	داده‌های ساختار یافته
پرس‌وجوها به زبان طبیعی	پرس‌وجوهای ساختار یافته
امکان بازیابی نتایج نامرتبط	عدم بازیابی نتایج غلط

برای پاسخگویی به نیاز اطلاعاتی کاربران، سیستم‌های بازیابی اطلاعات باید محتوای اطلاعات را تفسیر و بر طبق میزان مرتبط بودن با پرس‌وجوی کاربر رتبه بندی[2] کنند. این تفسیر شامل استخراج اطلاعات از اسناد به منظور انطباق آن‌ها با نیاز کاربران است. سختی کار نه تنها دانستن چگونگی استخراج اطلاعات می‌باشد بلکه دانستن چگونگی تصمیم‌گیری در مورد مرتبط بودن اسناد نیز می‌باشد. هدف اصلی بازیابی اطلاعات، بازیابی تمامی اسنادی است که با پرس‌وجوی کاربر مرتبط می‌باشند و امکان بازیابی تعداد کمی اسناد نامرتبط نیز وجود دارد.

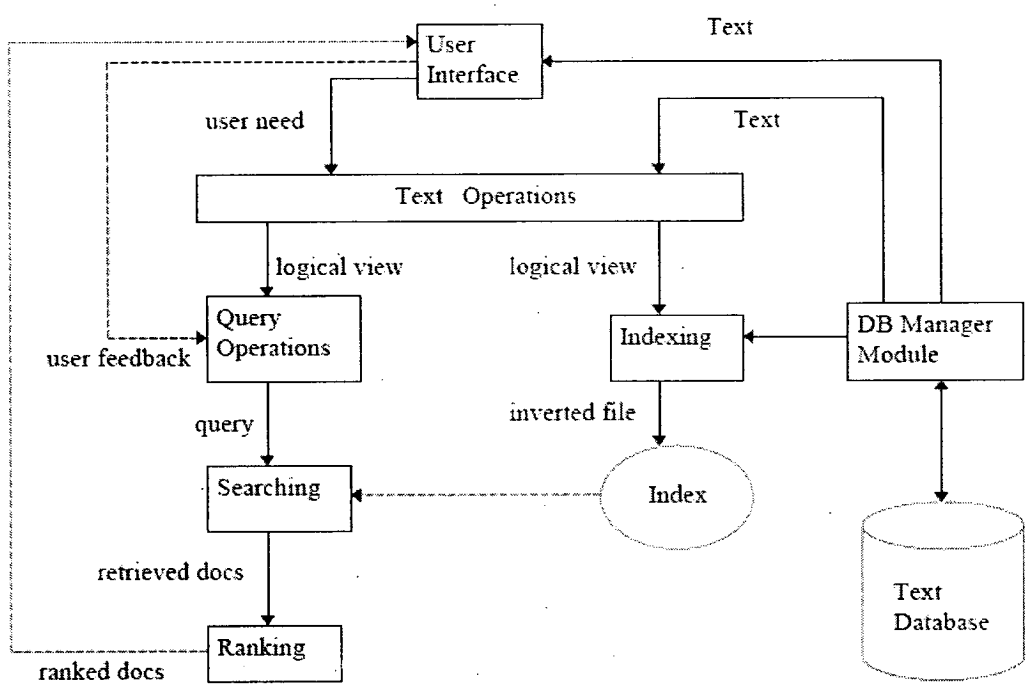
۱-۲-۲ فرآیند بازیابی اطلاعات

در شکل ۱-۱ فرآیند بازیابی اطلاعات نشان داده شده است. به منظور فراهم سازی امکان بازیابی، لازم است که پایگاه متون تعریف شود. این کار با استفاده از مدیر پایگاه داده انجام می‌شود، مدیر پایگاه داده موارد زیر را

---

[1] Relational Database
[2] Rank

مشخص می‌کند: الف- اسنادی که استفاده می‌شوند. ب- اعمالی که روی متون انجام می‌شود. ج- مدل متن (ساختار متن و عناصری که می‌توانند بازیابی شوند). اعمال متنی اسناد را تغییر شکل می‌دهند و دیدی منطقی از آن‌ها تولید می‌کنند به این ترتیب که هر سند را با مجموعه‌ای از کلید واژه‌ها نشان می‌دهند. کلید واژه‌ها یا به صورت خودکار از اسناد استخراج می‌شوند و یا به صورت دستی برای هر سند مشخص می‌شوند. برای استخراج کلید واژه‌ها، کلمات توقف[1] از متن حذف می‌شوند. کلمات توقف کلماتی هستند که تقریبا در تمامی اسناد رخ می‌دهند و قادر به تفکیک اسناد مرتبط از اسناد نامرتبط نمی‌باشند. بقیه‌ی واژه‌ها ریشه‌یابی می‌شوند تا تعداد کلیدواژه‌ها کاهش پیدا کند، به این پردازش‌ها اعمال متنی گفته می‌شود.



شکل ۱- ۱ فرآیند بازیابی اطلاعات [۲]

با ساخته شدن دید منطقی از اسناد، مدیر پایگاه داده، ایندکسی از متون می‌سازد. شاخص ساختار مهمی است و امکان جستجوی سریع روی حجم عظیمی از داده‌ها را فراهم می‌کند. معروف‌ترین ساختار شاخص فایل معکوس است. در فایل معکوس هر واژه به اسنادی اشاره می‌کند که این واژه در آن‌ها رخ داده است. اگرچه تولید شاخص

[1] Stopwords

و پایگاه داده هزینه‌ی زمانی و حافظه‌ای در بر دارد اما چون از سیستم‌های بازیابی اطلاعات بسیار زیاد استفاده می‌شود، این هزینه قابل چشم پوشی است.

پس از ایندکس گذاری اسناد پایگاه داده می‌توان عمل جستجو را انجام داد. ابتدا کاربر نیاز اطلاعاتی خود را در قالب یک پرس‌وجو مشخص می‌کند. نیاز اطلاعاتی کاربر توسط همان اعمال متنی که به اسناد اعمال شده‌اند، مورد پردازش قرار می‌گیرد و تجزیه می‌شود. بعد از آن اعمال پرس‌وجو انجام می‌شود و نیاز کاربر به فرم قابل ارائه به سیستم تبدیل می‌شود. سپس پرس‌وجو برای استخراج اسناد پردازش می‌شود. با وجود شاخص، پرس‌وجو به سرعت پردازش می‌شود.

قبل از ارسال نتایج به کاربر اسناد بازیابی شده بر اساس میزان مرتبط بودن با نیاز اطلاعاتی کاربر رتبه بندی می‌شوند. سپس کاربر اسناد رتبه بندی شده را به منظور پیدا کردن اطلاعات مفید بررسی می‌کند. در این مرحله ممکن است، کاربر زیرمجموعه‌ای از اسنادی که مشاهده کرده است را به عنوان اسناد دلخواه مشخص کند و چرخه‌ی بازخورد کاربر را آغاز کند. در این چرخه، سیستم با استفاده از اسنادی که کاربر انتخاب کرده است، اطلاعات پرس‌وجو را تغییر می‌دهد و به احتمال زیاد این پرس‌وجوی تغییر یافته نمایش بهتری از نیاز واقعی کاربر فراهم می‌کند.

۱-۲-۳ بازیابی اطلاعات کلیدواژه‌ای[1]

در سیستم‌های بازیابی اطلاعات کلید واژه‌ای معیار میزان مرتبط بودن اسناد با پرس‌وجوی کاربر و رتبه بندی اسناد، تعداد رخداد واژه‌های پرس‌وجو در آن‌ها می‌باشد. در این سیستم‌ها هر سند و هر پرس‌وجو به عنوان مجموعه‌ای از واژه‌ها در نظر گرفته می‌شود و سندی که بیشترین تعداد رخداد واژه‌های پرس‌وجو را داشته باشد، به عنوان مرتبط‌ترین منبع برای کاربر محسوب می‌شود.

اگرچه موثر بودن بازیابی کلیدواژه‌ای ثابت شده است اما این نوع بازیابی اطلاعات با مشکلاتی نیز مواجه است، این مشکلات را می‌توان به سه دسته‌ی زیر تقسیم‌بندی کرد:

۱- مشکلاتی که از ذات زبان طبیعی ناشی می‌شوند:

---

[1] Keyword

- عدم تطابق[1] واژگان. مشکل عدم تطابق واژگان به این علت رخ می‌دهد که مفاهیم ممکن است با واژگان متفاوتی در پرس‌وجوی کاربر و اسناد بیان شوند [۳]. در هر زبانی ممکن است که واژه‌ای یک یا چندین واژه‌ی هم معنا[2] داشته باشد. بنابراین، این احتمال وجود دارد که کاربران برای توصیف مفاهیم در پرس‌وجوهای خود از واژگانی متفاوت با واژگانی که نویسندگان برای توصیف همان مفهوم به کار می‌برند، استفاده کنند. علاوه براین نویسندگان متفاوت از واژگان متفاوتی برای بیان مفاهیم یکسان استفاده می‌کنند [٤]. برطبق [٥] کاربران با نیاز اطلاعاتی یکسان تنها در ۲۰% موارد پرس‌وجوی یکسان به کار می‌برند، بنابراین با وجود نیاز یکسان نتایج متفاوتی دریافت می‌کنند. در این موتورهای جستجو مراجعی که هر چند برای کاربر بسیار مفیدند ولی به دلیل عدم وجود واژگان پرس‌وجوی کاربر در آن‌ها و رخداد واژگان هم‌معنا در نتایج جستجو دیده نمی‌شوند. اگر مشکل عدم تطابق در سیستم‌های بازیابی اطلاعات به طور مناسبی حل نشود، کارایی آن‌ها کاهش پیدا خواهد کرد.
- واژگان چندین معنایی[3]. علاوه بر وجود واژگان هم‌معنا، واژگانی نیز وجود دارند که چندین معنا دارند و معنای آن‌ها با توجه به متنی که در آن به کاربرده شده‌اند، مشخص می‌شود. واژگان چند معنایی مشکلاتی را برای موتورهای جستجو به وجود می‌آورند؛ وقتی کاربر در پرس و جوی خود واژه‌ای را به کار می‌برد، موتورهای جستجویی که تنها بر اساس انطباق واژگان عمل جستجو را انجام می‌دهند، تنها مراجعی را به کاربر ارائه می‌دهند که واژه‌ی مذکور در آن‌ها به کار رفته باشد. پس ممکن است نتایجی ارائه شود که واژه‌ی مذکور در معنایی غیر از معنای مورد نظر کاربر در آن‌ها استفاده شده باشد [٦].

بنابراین با وجود واژگان هم‌معنا و چندمعنا، در بازیابی اطلاعات کلیدواژه‌ای نتایج جستجو بسیار حساس به واژه‌های به کار رفته در پرس‌وجوی کاربر است، به این ترتیب که ممکن است جستجو با واژه‌ای نتیجه‌ای در بر نداشته باشد در حالی که با واژه‌ی مترادف آن به نتیجه برسد. علاوه براین صحت نتایج جستجو منوط به این است که واژه‌هایی که هم در پرس‌وجوی کاربر و هم در اسناد بازیابی شده وجود دارند، به یک معنا باشند.

---

[1] Mismatch
[2] Synonym
[3] Polysemy

۲- کاربران

- پرس‌وجوهای کوتاه. کاربران در اغلب موارد اطلاعات مورد نیاز خود را توسط چند واژه بیان می‌کنند، بنابراین معمولا پرس‌وجوهای کاربران کوتاه می‌باشند. در [۷] بیان شده است که وقتی تعداد کلمات پرس و جو زیاد باشد، گسترش دادن پرس‌وجو کارایی زیادی به دنبال نخواهد داشت ولی پرس‌وجوهای کوتاه معمولا از گسترش پرس‌وجو سود می‌برند. در واقع علت این مساله ابهام کم‌تر پرس‌وجوهای طولانی و احتمال بیشتر رخداد واژه‌های پرس‌وجوی طولانی در اسناد مرتبط با آن‌ها می‌باشد. بنابراین مساله‌ی عدم تطابق در مورد پرس‌وجوهای کوتاه نسبت به پرس‌وجوهای طولانی شدیدتر است. ذکر این نکته ضروری است که معمولا پرس‌وجوهای کاربران کوتاه می‌باشند، به طوری‌که متوسط طول پرس‌وجو های کاربران دو کلمه ذکر شده است [۸].
- دانش ناقص کاربران. کاربران در اغلب موارد، دانش کاملی در مورد موضوع مورد نظر خود ندارند، بنابراین پرس‌وجوهای کاربران معمولا مبهم‌اند و بیان ضعیفی از نیاز اطلاعاتی کاربر می‌باشند. پرس‌وجوهای ضعیف بازیابی ضعیف اطلاعات را به دنبال خواهند داشت.

۳- سیستم‌های بازیابی اطلاعات. علاوه بر مسائلی که تاکنون مطرح شد، در بازیابی اطلاعات کلیدواژه‌ای تعداد نتایج زیاد است، در بین آن‌ها موارد نامرتبط و تکراری زیاد دیده می‌شود و نتایج مرتبط زیادی نیز بازیابی نمی‌شوند. بنابراین بعد از ارائه‌ی نتایج، کاربر باید آن ها را بررسی کند و از بین آن ها موارد مرتبط را پیدا کند [۹، ۱۰]. علاوه براین، تعداد زیاد نتایج جستجو از یک طرف باعث سر در گمی کاربران می‌شود و از سوی دیگر باعث می‌شود کاربران تنها چند نتیجه‌ی اولیه‌ی بازیابی را بررسی کنند. در نتیجه ممکن است کاربر به سراغ سند مرتبطی که جزء آخرین نتایج بازیابی است، نرود.

برای روشن‌تر شدن مساله در این جا به ارائه‌ی یک مثال پرداخته می‌شود. پرس‌وجوی "Apple Growing" را در نظر بگیرید، همان طور که در شکل ۱-۲ نیز مشاهده می‌شود، در نتایجی که با موتور جستجوی گوگل بازیابی شده است، "Apple" به معنای سیب در نظر گرفته شده است. بنابراین اگر نیاز اطلاعاتی کاربر کامپیوترهای اپل باشد با این پرس‌وجو به نتیجه‌ای نمی‌رسد و همانطور که در شکل ۱-۳ نیز مشاهده می‌شود کاربر مجبور است پرس‌وجوی خود را تغییر دهد. با این مثال مشخص می‌شود که نتایج جستجوی کلید واژه‌ای حساس به واژه‌های به کار رفته در پرس‌وجوی کاربر است.