شناسایی و تحلیل واژگان عمومی در منابع وب: رویکردی نو به بسط عبارت جست‌وجو با استفاده از زبان طبیعی در موتورهای کاوش[1]

رحمت الله فتاحی[2]

چکیده

بازیابی اطلاعات دقیق و مرتبط همواره از اهمیت ویژه‌ای در پژوهش‌های حوزه بازیابی اطلاعات برخوردار بوده است. مقاله حاضر، رویکرد جدیدی را در این زمینه معرفی و تجربه کرده است. بسط جست‌وجو با استفاده از واژه‌های عمومی که همراه کلیدواژه‌های موضوعی در منابع و متون موجود در وب ظاهر می‌شوند می‌تواند موجب افزایش میزان دقت و ربط در نتایج بازیابی در موتورهای کاوش شود. به این منظور، پژوهشی در دو مرحله انجام گرفت. در مرحله نخست، تعداد ۸۰۰ صفحه وب با استفاده از روش تحلیل محتوا مورد بررسی قرار گرفت و واژه‌های عمومی همراه هر کلیدواژه موضوعی شناسایی شد (جمعا ۴۲۶۴ مورد). نتایج تحلیل آماری نشان داد که ۱۴/۵٪ از واژه‌های عمومی میان دو حوزه مشترک هستند، یعنی عمومی مطلق به‌شمار می‌آیند، ۸۵/۵٪ واژه‌های عمومی وابسته به حوزه موضوعی می‌باشند، ۶۵٪ درصد واژه‌های عمومی پیش و ۳۵٪ از آنها پس از

---

۱. این مقاله در همایش انجمن کتابداری و اطلاع‌رسانی ایران که در تاریخ ۸۵/۳/۳ در **تالار یوسف شریعت‌زاده کتابخانه ملی ایران** برگزار گردید، ارائه شد.

۲. استاد دانشگاه فردوسی مشهد: fattahi@ferdowsi.um.ac.ir

کلیدواژه‌های موضوعی در متون ظاهر می‌شوند. همچنین، از نظر نوع واژه‌ها، ۶۲/۴٪ غیرموضوعی و ۳۷/۶٪ نیمه موضوعی به‌شمار می‌روند.
در مرحله دوم پژوهش، با اضافه کردن واژه‌های عمومی به کلیدواژه‌های موضوعی اولیه و انجام دوباره جست‌وجو (یعنی بسط جست‌وجو) در ۴ حالت جست‌وجوی کلیدواژه‌ای، عبارت دقیق، عنوان دقیق، و نشانی اینترنتی دقیق در گوگل، مشخص شد که این کار موجب بازیابی نتایج بسیار دقیق‌تر و مرتبط‌تر می‌شود. نتایج نشان داد که تفاوت معناداری در نتایج جست‌وجوی کلیدواژه‌ای و عبارتی دقیق میان دو حوزه وجود دارد. همچنین، میان دو حوزه از نظر نتایج جست‌وجوی دقیق عنوانی و نشانی اینترنتی، تفاوت معنادار مشاهده شد. همچنین، آزمون نسبت بسامد نتایج بازیابی در حالت جست‌وجوی کلیدواژه‌ای به سایر حالت‌ها نشان داد که تفاوت معناداری میان دو حوزه وجود دارد. در مجموع، نتایج پژوهش نشان داد که بسط جست‌وجو در حالت جست‌وجوی عنوانی و نشانی اینترنتی در هر حوزه موضوعی می‌تواند بسیار موفقیت‌آمیز باشد. به‌همین جهت، پیشنهاد می‌شود که موتورهای جست‌وجو، پیش‌فرض جست‌وجو را به دو حالت عنوانی و نشانی اینترنتی محدود کنند زیرا نتایج دقیق، مرتبط و کافی بازیابی خواهد شد. همچنین، می‌توان یک سیاهه از واژگان عمومی ایجاد کرده و به‌منزله یک ابزار هوشمند در موتورهای کاوش تعبیه کرد تا در بسط جست‌وجو مورد استفاده کاربران قرار گیرد.

کلیدواژه‌ها: بسط جست‌وجو/ بازیابی اطلاعات در وب/ زبان جست‌وجو/ واژگان عمومی/ موتورهای کاوش

مقدمه: مروری بر برخی مسائل و مشکلات بازیابی اطلاعات در وب

اکنون شبکه وب به‌صورت یکی از مهم‌ترین منابع دسترسی به اطلاعات، چه عمومی و چه علمی درآمده است. با این حال، بازیابی اطلاعات مرتبط در وب و شناسایی مفیدترین سایت‌ها و منابع برای بسیاری از کاربران دشوار است. به‌طور معمول، تعداد سایت‌های بازیابی شده در پاسخ به جست‌وجو بسیار زیاد است و

کاربران قادر نیستند وقت فراوانی را صرف ملاحظه صفحات بازیابی شده، کنند. در این راستا، یکی از دشواری‌های مطرح برای اغلب کاربران آن است که چگونگی جست‌وجوی مطالب مورد نظر خود را نمی‌دانند. برای مثال، چنانچه دانشجویی بخواهد اطلاعات ساده و کم حجمی را درباره مفهوم «جهانی شدن» بازیابی کند، معمولاً از واژه «جهانی شدن» به‌تنهایی استفاده می‌کند. این امر موجب بازیابی بیش از چند میلیون صفحه وب خواهد شد و وی قادر نخواهد بود در میان آنها، مطالب ساده و کم حجم مورد نظر خود را بیابد.

موتورهای کاوش متعدد و هوشمندی وجود دارد که از قابلیت‌های خوبی برای جست‌وجوی اطلاعات برخوردارند. بسیاری از کاربران نیز از کار با این ابزارها راضی هستند و به‌هر صورت از اطلاعات بازیابی شده استفاده می‌کنند. با این‌حال، درصد قابل ملاحظه‌ای از کاربران از نتایج بازیابی خرسند نبوده و آنها را بی‌ربط یا کم‌ربط می‌یابند. صاحب‌نظران و پژوهشگران متعددی بر این نکته به‌منزله نقطه ضعف موتورهای کاوش تاکید کرده‌اند. (Casasola and Gauch: 1997; Sugiura and Etzioni: 2000; Chowdhury and Soboroff: 2002; Soboroff: 2004; Pokorny: 2004).

برخی از مشکلات عمده بازیابی اطلاعات در موتورهای کاوش عبارتند از :

۱. انتخاب واژه‌های مناسب برای جست‌وجوی مطلوب برای بسیاری از کاربران دشوار است (Chowdhury: 1999; Chowdhury and Chowdhury: 1999; and Voorbij: 1999; Filman and Pant 1998; Lawrence and Giles 1998; Doan et al.: 1999; Lykke and Ingwersen: 1999; Baeza-Yates: Hurtado and Mendoza: 2004). هم‌چنین، اغلب کاربران معمولی با شیوه انجام جست‌وجوی عبارتی[1] بیگانه‌اند؛ زیرا چگونگی انتخاب و ترکیب واژه‌ها را به‌منظور تدوین عبارت مورد نظر نمی‌دانند.

۲. جست‌وجوی کلیدواژه‌ای که به‌منزله پیش‌فرض[2] در اغلب موتورهای کاوش در نظر گرفته شده موجب افزایش بسیار حجیم نتایج بازیابی[1] و در عین‌حال پایین

1. Phrase searching
2. Default

آمدن میزان دقت[2] می‌شود. (Sugiura and Etzioni: 2000; Tillett: 2001; Chowdhury and Soboroff: 2002; Soboroff: 2004).

۳. مرور سایت‌ها و منابع بازیابی شده بسیار وقت‌گیر و گاهی اوقات آزاردهنده است. (Oyama: 1999; Dias et al.: 1999; Spink and Xu: 2000).

۴. فهم اینکه موتورهای کاوش چگونه کار می‌کنند و چگونه واژه‌ها یا عبارت جست‌وجو را تفسیر می‌کنند برای اغلب کاربران دشوار است. به‌همین دلیل، بسیاری از کاربران از ادامه کار و اصلاح عبارت جست‌وجو صرف‌نظر می‌کنند. (Spink et al.: 1998; Ellis: Ford: & Furner: 1998). کاربران معمولاً با همان حداقل واژه‌هایی که در ذهن دارند اقدام به جست‌وجو در موتورهای کاوش می‌کنند. به بیان دیگر، پالایش و یا بسط جست‌وجو مورد توجه آنها قرار نمی‌گیرد.

۵. بسط و یا پالایش جست‌وجو برای بسیاری از کاربران ناآشنا و مبهم است و یا آنها ترجیح می‌دهند این گزینه را به‌دلیل دشواری در فهم و کاربرد آن نادیده بگیرند (Jansen: Spink and Saracevic: 2000). کاربران معمولاً با همان کلیدواژه‌هایی که به ذهن آنها می‌رسد جست‌وجو را آغاز می‌کنند و برای ادامه کار، کمتر از واژه‌های جدید استفاده می‌کنند. (Lykke and Ingwersen: 1999).

۶. امکانات موتورهای کاوش نیز برای کمک به کاربران برای محدود کردن و یا بسط جست‌وجو کافی نیست. تنها تعداد بسیار محدودی از موتورهای کاوش از قابلیت لازم برای بسط جست‌وجو از طریق پیشنهاد واژه‌ها یا عبارت‌های لازم برخوردارند (Baeza-Yates: Hurtado and Mendoza: 2004).

نکته اصلی در مورد مسائل مرتبط با بازیابی اطلاعات در وب آن است که، هنگام جست‌وجو، کاربران به‌طور معمول از واژه‌هایی استفاده می‌کنند که فقط جنبه‌های موضوعی نیاز اطلاعاتی آنها را نشان می‌دهد. چنانچه آنها جست‌وجوی خود را با واژه‌های دیگری برای یافتن اطلاعات ادامه می‌دهند که آن واژه‌ها نیز

---

1. Recall
2. Precision

دال بر موضوع هستند می‌توانند به اطلاعات بیشتری دست پیدا کنند Furnas et al.: 1987 and Iivonen: 1995). به بیان دیگر، در حالی‌که ممکن است جنبه‌های دیگری در کنار موضوع مربوطه مورد نظر باشد، آنها قادر نیستند واژه‌ها یا عبارت‌های بیانگر جنبه‌های دیگر را از ذهن خود فراخوانی کنند. چنانچه به مثال "جهانی شدن" برگردیم، واژه‌هایی را می‌توان به کلیدواژه "جهانی شدن" افزود که دال بر جنبه‌های خاص دیگری هستند و می‌توانند سایت‌ها و منابعی را بازیابی کنند که مفهوم جهانی شدن را به زبانی ساده بیان کرده باشند. برای مثال عبارت‌هایی چون "آشنایی با جهانی شدن"، "مقدمه‌ای بر جهانی شدن"، "درباره جهانی شدن"، "جهانی شدن چیست؟"، و "تعریف جهانی شدن" قادرند جنبه ساده بودن و مقدماتی بودن موضوع را بیان کنند. این گونه عبارت‌ها و واژه‌ها را می‌توان "واژه‌های غیرموضوعی[1]" یا "واژه‌های عمومی[2]" نامید؛ واژه‌هایی که می‌توانند در بسط جست‌وجو و بازیابی منابع دلخواه مورد استفاده قرار گیرند. با توضیحات بالا، می‌توان هدف پژوهش حاضر را شناسایی واژه‌های غیرموضوعی و بررسی میزان قابلیت آنها در خاص کردن محدوده جست‌وجو (بسط جست‌وجو) بیان کرد.

تعریف مفاهیم کلیدی پژوهش

بسط جست‌وجو: فرایند پالایش عبارت جست‌وجو به‌منظور جلوگیری از بازیابی بسیار زیاد و یا کم ربط. این عمل با افزودن یک یا چند واژه به‌عبارت اولیه جست‌وجو انجام می‌شود و باعث می‌گردد که کاربر بتواند نیاز اطلاعاتی خود را از طریق این واژه‌های کمکی به‌شکل بهتری بیان کند. به بیان دیگر، بسط جست‌وجو به‌منظور تدوین عبارت جست‌وجو به‌شکل مناسب‌تر انجام می‌شود و هدف آن افزایش میزان "دقت" در نتایج بازیابی است به‌صورتی که منابع بازیابی

1. Non-topical
2. General

شده جنبه‌ها یا ویژگی مورد نظر جست‌وجوگر را در مورد منابع اطلاعاتی دربرداشته باشد، جنبه‌هایی مانند سطح مطلب، مخاطبان مورد نظر، عمق موضوع، رویکرد یا نوع منبع.

واژه‌های موضوعی[1]: این واژه‌ها، محتوای موضوعی منابع اطلاعاتی را نشان می‌دهند. اغلب کاربران نیاز اطلاعاتی خود را در قالب واژه‌های موضوعی بیان می‌کنند و به موتورهای کاوش می‌سپارند. برای مثال "جهانی شدن"، "تروریسم"، "سرطان پوست"، و "بهداشت روانی" واژه‌های موضوعی به‌شمار می‌آیند.

واژه‌های غیرموضوعی[2]: اینها واژه‌های عمومی هستند که معمولاً به‌تنهایی مورد جست‌وجو قرار نمی‌گیرند زیرا به‌خودی‌خود معنا و مفهوم خاصی ندارند. واژه‌های عمومی همواره به همراه (پیش یا پس از) واژه‌های موضوعی می‌آیند تا جنبه خاصی از آن موضوع را نشان دهند. برای مثال؛ "مقدمه‌ای بر ..."، "آشنایی با ..."، "درباره ..."، "تاریخ ..." و مواردی از این قبیل، غیرموضوعی یا عمومی تلقی می‌شوند.

واژه‌های نیمه‌موضوعی[3]: این واژه‌ها نیز به‌طور معمول به‌تنهایی مورد جست‌وجو قرار نمی‌گیرند بلکه مانند واژه‌های غیرموضوعی همراه کلیدواژه‌های موضوعی می‌آیند، مانند "ریسک ...""پیشگیری از ..."، "حادثه ..."و مانند آنها.

لازم به اشاره است که به‌لحاظ معناشناختی، واژه‌های غیرموضوعی از جایگاه خاصی برخوردارند زیرا به فراوانی در متون علمی و غیرعلمی مورد استفاده قرار می‌گیرند. بنابراین، جزء جدایی‌ناپذیر زبان به‌شمار می‌روند و نقش ویژه‌ای در بیان رویکرد، سطح خوانایی، نوع مطلب یا منبع و غیره دارند. هم چنانکه واژه‌های مرتبط اعم، اخص، و یا مترادف در متون و هم چنین در اصطلاحنامه‌ها (تزاروس‌ها) به‌لحاظ معناشناختی و در چارچوب هستی‌شناسی[4] واژگان نقش

---

1. Topical terms
2. Non-topical terms
3. Semi-topical terms
4. Ontology

دارند، واژه‌های عمومی هم در بیان مقاصد پدیدآورندگان متـون و تولیدکننـدگان دانش اهمیت دارند. در واقع و از دیدگاهی دیگر، واژه‌های غیرموضوعی (عمـومی) را می‌توان زیرمجموعه مرتبط با واژه‌های موضوعی به‌شمار آورد.

مرور پیشینه پژوهش

موضوع "بسط جست‌وجو" و زمینه‌های مرتبط با آن از مدت‌ها پیش و بـه‌مـوازات توسعه نظام‌های بازیابی اطلاعات بـه‌ویـژه پـس از ابـداع شـبکه وب مـورد توجـه پژوهشگران قرار گرفته است. به‌همین جهت، ادبیات پژوهشی این حوزه به‌ویژه در زمینه چگونگی تدوین عبارت جست‌وجو و افزایش دقـت و ربـط بـسیار گـسترده است (برای مثال، نگاه کنید به: Harman: 1988; Anick and Tipirneni: 1999; Efthimiadis: 1995: 2000; McArthur and Bruza: 2000; Bruza: McArthur and Dennis: 2000; Billerbeck and Zobel: 2003).[1]

هدف عمده این‌گونه پژوهش‌هـا شناسـایی رفتـار جـست‌وجـوگران در تـدوین عبارت جست‌وجـو در موتورهـای کـاوش و اینکـه ایـن ابزارهـا چگونـه بـه بـسط جست‌وجو کمک می‌کنند می‌باشـد. توجـه اغلـب پـژوهش‌هـا بـر جـست‌وجـوی موضوعی و استفاده از فنون دسته‌بندی موضوعی[2] به‌منظور کسب نتایج بهتر است. در این راستا، کاربرد واژه‌نامه‌ها به‌منزله یک رویکـرد نـو در بـسط جـست‌وجـوی موضوعی و افزایش ربط در نتایج بازیابی، توجه برخی از پژوهـشگران را بـه خـود جلب کرده است. در عین‌حال، نکته‌ای که از سوی اغلب پژوهشگران مورد غفلـت قرار گرفته است آن است که جست‌وجوی اصطلاحنامه‌ای تنها واژه‌های اعم، اخص و مرتبط را در دسترسی کاربران قرار می‌دهد و جنبه‌هـای مـرتبط بـا موضـوعات مورد جست‌وجو، مانند مواردی که پیشتر برشمرده شد (سطح و رویکـرد مطلـب، نوع منبع، . . . .) به آنها نمایانده نمی‌شود. به بیان دیگر، استفاده از اصـطلاحنامه

---

۱. سیاهه بسیاری از منابع پژوهشی در این زمینه را می‌توانید در سایت زیر بیابید:
http://wolton.liu.edu/docis/

2. Subject Clustering Techniques

برای افزایش دقت در نتایج بازیابی تنها به جنبه‌های موضوعی منحصر شده در حالی‌که امکان دارد بسیاری از کاربران از جنبه خاص دیگری در جست‌وجوی اطلاعات باشند.

مروری بر ادبیات مرتبط با بسط جست‌وجو نشان می‌دهد که پژوهش‌های بسیار کمی در زمینه کاربرد واژه‌های عمومی (غیرموضوعی) انجام شده است. یکی از پژوهش‌هایی که تا اندازه ای همسو با پژوهش حاضر بوده است توسط سوگیورا و اتزیونی (Sugiura and Etzioni: 2000) صورت گرفته که طی آن امکان استفاده از واژه‌های عمومی برای مسیریابی جست‌وجو مورد بررسی قرار گرفت. آنها با ایجاد یک ابزار مسیریابی[1] به نام Q-Pilot به این نتیجه رسیدند که عبارت‌های جست‌وجویی را که حاوی واژه‌های تخصصی و واژه‌های عمومی است می‌توان به موتورهای تخصصی مربوطه هدایت کرد. این کار از طریق استخراج و دسته‌بندی واژه‌ها در دو گروه موضوعی و غیرموضوعی انجام می‌شد. تاکید پژوهش سوگیورا و اتزیونی بر معماری ابزار مسیریابی عبارت جست‌وجو و هدایت آن به‌سوی موتورهای تخصصی مربوطه بود. به‌همین جهت، تلاشی برای شناسایی و دسته‌بندی اساسی واژه‌های عمومی انجام ندادند. همچنین، اساس کار آنها بر استفاده از راهبرد جست‌وجوی بولی با استفاده از "AND" بود و نه بر کاربرد واژگان عمومی در قالب عبارت‌های زبان طبیعی.

پژوهش دیگری که در زمینه استفاده از واژه‌های غیرموضوعی برای بسط جست‌وجو صورت گرفته است توسط چان (Chan: 2000) انجام شد. وی با استفاده از اصطلاح‌های سرعنوان‌های موضوعی کتابخانه کنگره (LCSH)، واژه‌های موضوعی و عمومی را به‌صورت جداگانه در قالب فراداده دوبلین کور جای داد تا از آنها بتوان برای جست‌وجو استفاده کرد. در آن پژوهش، واژه‌های عمومی فقط شامل تقسیمات فرعی (مانند جغرافیایی، تاریخی و شکلی) موجود در LCSH می‌شد و سایر واژه‌های عمومی را که عملا به زبان طبیعی در متون به کار برده

1. Query Routing System

می‌شوند دربر نداشت. به‌همین جهت در این رویکرد، امکان لحاظ کردن جنبه‌های خاص در بسط جست‌وجو بسیار محدود بود. تنها رویکردی که با پژوهش حاضر مشابه می‌باشد توسط موتور کاوش AskJeeves[1] به‌کارگرفته شده است. این موتور کاوش براساس واژگان منابع اینترنتی، ابزاری با عنوان ZOOM برای پالایش جست‌وجو ایجاد کرده که به‌موازات جست‌وجوی کلیدواژه‌ای توسط کاربر، بسط جست‌وجو نیز به‌طور خودکار و با استفاده از واژه‌های عمومی صورت می‌گیرد و نتایج بازیابی در سمت راست صفحه نمایش به‌صورت یک سیاهه شامل کلیدواژه‌های موضوعی و واژه‌های عمومی همراه آنها نشان داده می‌شود. با استفاده از این سیاهه، کاربران می‌توانند روی گزینه دلخواه که جنبه مورد نظر آنها را نشان می‌دهد کلیک کنند تا بسط جست‌وجو روی آنها صورت گیرد. آنچه در مورد این رویکرد می‌توان مشاهده کرد آن است که، با وجود آنکه از ساختار فنی نسبتا مطلوبی برخوردار است، اما دامنه واژه‌های عمومی در سیاهه ZOOM محدود بوده و در برخی موارد کاستی‌هایی در ارتباط میان کلیدواژه‌های موضوعی و واژه‌های عمومی مشاهده می‌شود. در برخی موارد نیز سایت یا صفحه مشخصی برای برخی عبارت‌های موجود در سیاهه وجود ندارد وآن عبارت‌ها مانند ارجاع کور به‌شمار می‌روند.

هدف، دامنه و رویکرد پژوهش حاضر با رویکرد موتور AskJeeves از چند جنبه متفاوت است: ۱) شناسایی واژه‌های عمومی که همراه کلیدواژه‌های موضوعی در متون وبی ظاهر می‌شوند، ۲) دسته‌بندی آن واژه‌ها از نظر میزان وابستگی حوزه‌ای[2]، ۳) تحلیل واژه‌های عمومی از نظر اینکه اغلب پیش از کلیدواژه‌های موضوعی ظاهر می‌شوند یا پس از آنها، و ۴) بررسی قابلیت این واژه‌ها از نظر بازیابی دقیق‌تر و مرتبط اطلاعاتی در ۴ حالت جست‌وجو

1. www.ask.com
2. Domain-specific terms

(جست‌وجوی کلیدواژه‌ای عمومی[1]، عبارتی دقیق[2]، عنوان دقیق[3]، و نشانی اینترنتی دقیق[4]) در موتور کاوش گوگل.

پرسش‌های پژوهش

در راستای هدف‌های مورد نظر پژوهش، پرسش‌های زیر مورد توجه قرار گرفت :

۱. پربسامدترین واژه‌های غیرموضوعی که همراه واژه‌های موضوعی در حوزه‌های: ۱) پزشکی و بهداشت[5]، و ۲) علوم اجتماعی[6] ظاهر می‌شوند کدام است؟

۲. پربسامدترین واژه‌های غیرموضوعی که پیش از و نیز پس از واژه‌های موضوعی می‌آیند کدام است؟

۳. پربسامدترین واژه‌های غیرموضوعی مشترک میان دو حوزه پزشکی و علوم اجتماعی کدام است؟

۴. آیا تفاوت معناداری میان منابع وبی حوزه‌های پزشکی و علوم اجتماعی از نظر بسامد واژه‌های غیرموضوعی وجود دارد؟

۵. بسامد منابع وبی بازیابی شده در انواع حالت‌های جست‌وجو (کلیدواژه‌ای، عبارتی دقیق، عنوان دقیق، و نشانی اینترنتی دقیق) چگونه است؟

روش و مراحل انجام پژوهش

در این پژوهش از روش تحلیل متن[7] که گونه‌ای از روش تحلیل محتوا است استفاده شد. تحلیل متن از طریق شمارش واژه‌ها یا عبارت‌ها در متن یا عبارت‌ها در متون موردنظر، شناسایی ارتباط میان واژه‌ها و نیز دسته‌بندی کردن آنها با

---

1. General keyword search
2. Exact phrase search
3. Exact title search
4. Exact URL search
5. Health
6. Social Sciences
7. Text analysis

توجه به هدف‌های پژوهش انجام می‌گیرد. تمرکز تحلیل متن، همچنین، می‌تواند با توجه به موقعیت و مکان واژه‌ها نسبت به هم در یک یا چند متن باشد. برای تحلیل متن، به‌طور معمول، از دو تکنیک "تحلیل عناصر[۱]" و"تحلیل ساختار[۲]" استفاده می‌شود، که اولی به‌منظور شناسایی واژه‌ها، گروه‌های واژه‌ها و بسامد آنها و مفهوم دوم به شناسایی ارتباط میان واژه‌ها انجام می‌گیرد (Hicks: Rush and Strong: 1977: p. 90).

پژوهش حاضر در دو مرحله انجام گرفت:

۱. مرحله اول: در این مرحله، ۱۰ کلیدواژه از حوزه پزشکی و بهداشت و ۱۰ کلیدواژه از حوزه علوم اجتماعی که به‌نظر می‌رسید از موضوعات روز و مورد علاقه کاربران باشند برگزیده و در گوگل جست‌وجو شد[۳]. موتور کاوش گوگل بدان جهت انتخاب شد که یکی از پراستفاده‌ترین ابزارها برای جست‌وجو در وب به‌شمار می‌رود (Griffith and Brophy: 2005). از نتایج بازیابی شده که در پاسخ به هر کدام از ۲۰ کلیدواژه موضوعی حاصل شد، ۱۰ وب‌سایت اول برگزیده شد. در مجموع، ۲۰۰ وب‌سایت مورد بررسی قرار گرفت و از هر یک از آنها ۴ صفحه نخست برای تحلیل متن پرینت گرفته شد. این امر با توجه به نتایج بسیاری از پژوهش‌ها صورت گرفت که نشان می‌داد اغلب کاربران تنها چند صفحه نخست هر سایت را مورد بررسی قرار می‌دهند (Spink: Greisdof and Bateman: 1998; Jansen: Spink and Saracevic: 2000; Henzinger: Motwani: Silverstein: 2002). به‌عبارت دیگر، در مجموع ۸۰۰ صفحه برای شناسایی واژه‌ها یا عبارت‌های عمومی که معمولاً پیش یا پس از کلیدواژه‌های موضوعی می‌آیند تحلیل شد. بدین ترتیب، واژه‌ها و عبارت‌های عمومی برگرفته شده وارد

---

1. Elemental analysis
2. Structural analysis

۳. کلیدواژه‌ها می‌تواند هر کلیدواژه‌ای باشد زیرا آنچه مهم است شناسایی واژه‌های عمومی موجود در متون است. ابتدا تصمیم آن بود که از واژه‌های پربسامدی که کاربران عملاً در موتورهای کاوش جستجو کرده‌اند استفاده شود اما ملاحظه شد که اغلب این واژه‌ها مربوط به هنرپیشه‌ها، خوانندگان و یا واژه‌های غیراخلاقی بود.

نرم‌افزار آماری spss شد تا: ۱) بسامد هر یک و مجموع آنها در دو حوزه مشخص شود، ۲) بسامد آنها با توجه به مکان آنها (پیش یا پس از کلیدواژه موضوعی) معلوم گردد، و ۳) بسامد واژه‌های عمومی مشترک میان دو حوزه نیز محاسبه شود. در مجموع، ۱۴۴۲ واژه و عبارت موضوعی شناسایی که پس از کنار گذاشتن واژه‌هایی که بسامد آنها زیر ۳ بود، در نهایت ۱۰۷۱ مورد برگزیده شد. تحلیل‌های فوق در مورد آنان انجام گرفت که یافته‌ها در جدول شماره ۱ ارائه شده است.

۲. مرحله دوم: این مرحله در سپتامبر ۲۰۰۵ به اجرا درآمد و طی آن ۲۰ کلیدواژه که در مرحله اول به‌صورت مستقل جست‌وجو شده بود، این بار با افزودن واژه‌های عمومی مرتبط با هر یک از آنها مجددا در ۴ حالت کاوش در گوگل مورد جست‌وجو قرار گرفت. در مجموع، ۴۲۹۲ جست‌وجو (۱۰۷۱ واژه عمومی در ۴ حالت کاوش) به اجرا درآمد. و بسامد بازیافت‌های مربوطه در حالت‌های ۴ گانه در نرم‌افزار آماری SPSS وارد شد. آنگاه تحلیل‌های آماری مانند آزمون t، آزمون مجذور کای، نسبت و درصد، روی داده‌ها انجام گرفت.

یافته‌های عمده پژوهش عبارتند از:

۱. پربسامدترین واژه‌ها و عبارت‌های عمومی در دو حوزه پزشکی و علوم اجتماعی

در پاسخ به این پرسش، ۱۰۷۱واژه عمومی (شناسایی شده در مرحله اول پژوهش) با توجه به بسامد آنها در ۸۰۰ صفحه مورد بررسی، مرتب شد. پس از این بررسی، ۵۰ واژه و عبارت پربسامدتر شناسایی شد (در پیوست شماره یک ارائه شده است). بسامد واژه‌ها از نظر تنوع (غیرموضوعی و موضوع وابسته[1])، مکان حضور (پیش یا پس از کلیدواژه موضوعی)، موارد مشترک میان دو حوزه (همپوشانی‌ها) در جدول شماره ۱ نشان داده شده است.

1. Domain specific

جدول ۱. تحلیل واژه‌های عمومی از نظر نوع، مکان و همپوشانی در حوزه پزشکی و علوم اجتماعی.

حوزه موضوعی	بسامد کل		مکان نسبت به واژه موضوعی				نوع			
			پیش از		پس از		غیرموضوعی		نیمه موضوعی	
	بسامد	%	بسامد	%	بسامد	%	بسامد	%	بسامد	%
پزشکی	387	36.13	232	59.9	155	40.1	235	60.72	152	39.28
علوم اجتماعی	684	63.87	464	67.8	220	32.2	434	63.46	250	36.54
جمع کل	1071	100	696	63.85	375	36.15	669	62.09	403	37.91
موارد مشترک	156	14.56	86	55.7	70	44.30	136	86.07	22	13.93

۲. تفاوت میان پزشکی و علوم اجتماعی با توجه به بسامد واژه‌های عمومی

با توجه به واژه‌ها و عبارت‌های دارای بسامد بالای ۳، میانگین بسامد واژه‌های عمومی در حوزه پزشکی ۴/۸۲ و در علوم اجتماعی ۶/۱۷ بود. این یافته می‌تواند بیانگر این نکته باشد که، به‌طور معمول، تعداد واژه‌های عمومی در متون حوزه علوم اجتماعی بیش از پزشکی است. به‌عبارت دیگر، متون حوزه علوم اجتماعی را می‌توان با واژه‌های عمومی متنوع‌تری جست‌وجو کرد.

۳. پربسامدترین واژه‌های مشترک میان دو حوزه پزشکی و علوم اجتماعی

همان‌طورکه در پیوست شماره یک و نیز در جدول شماره ۱ آمده است، ۷۸ جفت (۱۵۶ مورد) واژه عمومی به‌صورت مشترک در دو حوزه وجود دارد. به‌عبارت دیگر، ۱۴/۷۱٪ همپوشانی میان واژه‌های عمومی دو حوزه وجود دارد که کاملا عمومی (عمومی مطلق) هستند و وابستگی به حوزه موضوعی ندارند (مانند about: introduction to: history of). در مقابل، ۸۵/۲۹٪ واژه‌های عمومی در حوزه(های) موضوعی خاصی کاربرد دارند (مانند Foundation: News: Tips). بر اساس این یافته می‌توان پیشنهاد کرد که واژگان عمومی خاص موتورهای کاوش تخصصی به‌صورت یک ابزار هوشمند و برای کمک به بسط جست‌وجو شناسایی شده و در آنها تعبیه شوند.

۴. مکان واژه‌ها و عبارت‌های عمومی

مکان واژه‌ها در عبارت جست‌وجو نقش مهمی در بسط جست‌وجو دارد. هر چه ترکیب واژه‌ها با هم و تشکیل عبارت جست‌وجو به زبان طبیعی و زبان متن نزدیکتر باشد، امکان بازیابی دقیق‌تر و مرتبط‌تر افزایش می‌یابد. یافته‌ها (جدول ۱) نشان می‌دهد که در حوزه پزشکی ۵۹/۹٪ واژه‌های عمومی و ۶۷/۸٪ واژه‌های عمومی در حوزه علوم اجتماعی پیش از کلیدواژه‌های موضوعی ظاهر می‌شوند. به‌عبارت دیگر، بیشتر واژه‌های عمومی به‌صورت طبیعی پیش از کلیدواژه‌های موضوعی) و در حالتی نزدیکتر به زبان طبیعی و زبان متون) می‌آیند. نتایج آزمون مجذور کای (۰/۰۰۱) نشان داد که تفاوت معناداری از نظر مکان واژه‌های عمومی میان دو حوزه وجود دارد بدان صورت که با اطمینان ۹۹٪ می‌توان گفت که بسامد واژه‌های عمومی پیش از کلیدواژه‌های موضوعی در حوزه علوم اجتماعی بیشتر از حوزه پزشکی است.

Ch-s	۱۶/۹۶۰
Df	۳
.Asymp. Sig	.۰۰۱

۵. بسامد نتایج بازیابی ناشی از بسط جست‌وجو در دو حوزه پزشکی و علوم اجتماعی

همان‌گونه که در بخش روش تحقیق توضیح داده شد، ۲۰ کلیدواژه موضوعی برگزیده شده، ابتدا به‌صورت مستقل و سپس با افزودن واژه‌های عمومی در ۴ حالت جست‌وجو در گوگل مورد جست‌وجو قرار گرفت.

بسامد و میانگین بسامد در جدول‌های زیر ارائه شده است.

جدول ۲. بسامد و میانگین بسامد صفحات پیش‌بینی شده در پاسخ به جست‌وجوی موضوعی در حوزه پزشکی

واژه‌ها/عبارات موضوعی	کلیدواژه‌ای	عبارتی دقیق	عنوان دقیق	نشانی دقیق
Suicide	۹۳۶۰۰۰۰۰	93:600:000	1:550:000	465:000
SARS	15:000:000	15:000:000	879:000	837:000
Occupational health	66:200:000	16:600:000	315:000	44:700
Eating disorder	9:510:000	2:090:000	85:000	41:500
Breast cancer	59:600:000	44:600:000	1:400:000	695:000
Skin care	54:600:000	13:800:000	2:470:000	879:000
Drug abuse	49:900:000	18:700:000	228:000	254:000
Cosmetic plastic surgery	10:500:000	525:000	32:300	2:550
Yoga	28:000:000	27:800:000	1:840:000	2:430:000
Mental health	142:000:000	93:500:000	1:660:000	2:340:000
میانگین	44:921:000	23:438:500	895:970	752:457

جدول ۳. بسامد و میانگین بسامد صفحات بازیابی شده در پاسخ به جست‌وجوی موضوعی در حوزه علوم اجتماعی

واژه‌ها/عبارات موضوعی	کلیدواژه‌ای	عبارتی دقیق	عنوان دقیق	نشانی دقیق
Child abuse	55:700:000	14:500:000	296:000	47:400
Abortion	12:500:000	12:600:000	360:000	164:000
Human cloning	13:900:000	1:770:000	50:400	819
Social security	273:000:000	110:000:000	1:840:000	384:000
Domestic violence	38:000:000	22:300:000	428:000	98:900
Globalization	51:000:000	51:000:000	563:000	287:000
Human rights	440:000:000	126:000:000	3:170:000	177:000
Terrorism	139:000:000	139:000:000	1:770:000	635:000
Adoption	127:000:000	127:000:000	1:660:000	2:300:000
Feminism	13:800:000	13:800:000	238:000	153:000
میانگین	116:390:000	61:797:000	1:037:540	424:712

همان‌گونه که مشاهده می‌شود، میانگین صفحات بازیابی شده در هر یک از ۴ حالت جست‌وجو در گوگل بسیار بیش از آنی است که کاربران بتوانند به بررسی آنها و یافتن منابع مرتبط بپردازند. حتی محدود کردن دامنه جست‌وجو از حالت "کلیدواژه‌ای عمومی" به "عبارتی دقیق"، به "عنوانی دقیق" و یا به "نشانی اینترنتی دقیق" باز هم دارای نتایجی با بسامد بسیار بالا است. صاحب‌نظران و پژوهشگران مختلف نیز (همچون Sugiura and Etzioni: 2000; Tillett: 2001; Chowdhury and Soboroff: 2002; Soboroff: 2004) بازیافت بسیار بالا را یک مسئله بسیار جدی برای کاربران می‌دانند. به بیان دیگر، انجام جست‌وجوهای کلی، بدون بسط آنها با واژه‌ها و عبارت‌های عمومی متناسب، نمی‌تواند به نتایج مطلوب منتهی شود. در مقابل، با افزودن واژه‌ها و عبارت‌های عمومی (غیرموضوعی) و بسط جست‌وجو، کاربران می‌توانند به نتایج کمتر اما دقیق‌تر و مطلوبتری برسند. یافته‌های پژوهش نیز موفقیت بسط جست‌وجو از طریق واژه‌های عمومی را نشان داد. جدول شماره ۴ نتیجه آزمون t را در زمینه تفاوت میانگین صفحات بازیابی شده در پاسخ به دو حالت پیش از بسط جست‌وجو و پس از بسط جست‌وجو در دو حوزه موضوعی نشان می‌دهد.

جدول ۴. آزمون t در مورد نتایج بازیابی و وجود تفاوت در دو حوزه پزشکی و علوم اجتماعی.

گزینه‌های جست‌وجو	حوزه موضوعی	تعداد واژه‌ها	میانگین بسامدها	انحراف استاندارد	|t|	df	P
کلیدواژه‌ای	پزشکی	387	9361203	13926204			
	علوم اجتماعی	683	20298788	33517754	6.126	1068	0.001
عبارتی دقیق	پزشکی	387	148476	690518			
	علوم اجتماعی	683	324080	1149376	2.738	1068	0.006
عنوان دقیق	پزشکی	387	3236	19175			
	علوم اجتماعی	683	3483	21496	0.188	1069	0.851
نشانی دقیق	پزشکی	387	754	8672			
	علوم اجتماعی	683	279	1940	1.374	1069	0.170

همان‌گونه که جدول فوق نشان می‌دهد، تفاوت معناداری (p=.001) میان دو حوزه از نظر میانگین بسامد نتایج جست‌وجوی کلیدواژه‌ای و جست‌وجوی عبارتی دقیق وجود دارد. به بیان دیگر، میانگین بازیافت‌ها در حوزه علوم اجتماعی در این دو حالت جست‌وجو بیشتر از حوزه پزشکی است. در عین‌حال، میان دو حوزه از نظر میانگین بازیافت‌های ناشی از بسط جست‌وجو در حالت جست‌وجوی عنوانی دقیق و نشانی دقیق تفاوتی وجود ندارد (p=. /851 و p=. /171) از این یافته می‌توان نتیجه گرفت که بسط جست‌وجو در دو حالت اخیر در هر دو حوزه به یک اندازه موفقیت‌آمیز است.

همچنین، برای نشان‌دادن ارزش بسط جست‌وجو با استفاده از واژه‌های عمومی، میانگین بسامد بازیافت‌های ناشی از بسط بر میانگین بسامد بازیافت‌های غیربسط یافته تقسیم شد و در چهار حالت جست‌وجو محاسبه گردید. آزمون نسبت و نتایج آن در زیر ارائه شده است.

جدول ۵. میانگین و نسبت بازیافت‌ها بیش و پس از بسط جست‌وجو در چهار حالت جست‌وجو در دو حوزه

$$\text{نسبت} = \frac{\text{میانگین بسامد بازیافت‌ها پس از بسط جست‌وجو}}{\text{میانگین بسامد بازیافت‌ها پیش از بسط جست‌وجو}} \times ۱۰۰$$

جدول ۵. میانگین و نسبت بازیافت‌ها بیش و پس از بسط جست‌وجو در چهار حالت جست‌وجو در دو حوزه

حوزه موضوعی	میانگین	کلیدواژه ای	عبارتی دقیق	عنوان دقیق	نشانی دقیق
پزشکی	پیش از بسط جست‌وجو	44:921:000	23:438:500	895:970	752:457
	پس از بسط جست‌وجو	9:332:489	148:094	3:227	752
	نسبت	%20.77	%0.63	%0.36	%0.1
علوم اجتماعی	پیش از بسط جست‌وجو	116:390:000	61:797:000	1:037:540	424:712
	پس از بسط جست‌وجو	20:258:444	382:811	3:505	279
	نسبت	%17.40	%0.61	%0.33	%0.06

نتایج آزمون t درمورد میانگین‌ها همچنین نشان داد که تفاوت معناداری از نظر نسبت میانگین‌ها برای دو حالت جست‌وجوی کلیدواژه‌ای و جست‌وجوی عبارتی دقیق

میان دو حوزه وجود دارد (جدول ۶). در عین‌حال، تفاوت معناداری میان نسبت‌ها در حالت‌های جست‌وجوی عنوانی دقیق و جست‌وجوی نشانی دقیق وجود ندارد. این یافته با یافته پیشین (جدول ۴) همخوانی دارد و نشان‌دهنده میزان تأثیر یکسان بسط جست‌وجو در حالت عنوان دقیق در دو حوزه پزشکی و علوم اجتماعی است. نسبت بازیافت صفحات وب را در دو حالت پیش از بسط و پس از بسط جست‌وجو در موتور کاوش می‌توان به‌گونه دیگر نیز محاسبه کرد. میانگین بازیافت‌های ناشی از بسط جست‌وجو را در سه شیوه جست‌وجوی "عبارت دقیق"، "عنوان دقیق"، و "نشانی دقیق" بر میانگین بازیافت در شیوه جست‌وجوی "کلیدواژه‌ای" تقسیم می‌کنیم تا میزان محدود شدن و کاهش نتایج بازیابی را در دو حوزه مقایسه کنیم و بدین ترتیب میزان اثر بسط جست‌وجو را مشخص کنیم. جدول‌های ۷ و ۸ یافته‌ها را نشان می‌دهد.

جدول ۷ :نسبت بازیافت‌ها در شیوه‌های گوناگون جست‌وجو به جست‌وجوی کلیدواژه‌ای در دو حوزه

حوزه وضوعی	حالت جست‌وجو	عبارتی دقیق به کلیدواژه‌ای	عنوان دقیق به کلیدواژه	شانی دقیق ه کلیدواژه
زشکی	پیش از بسط جست‌وجو	%52	%1.99	%1.67
	پس از بسط جست‌وجو	%1.58	%0.34	%0.008
علوم جتماعی	پیش از بسط جست‌وجو	%52	%0.89	%0.36
	پس از بسط جست‌وجو	%1.88	%0.017	%0.001

جدول ۸. آزمون t برای میانگین‌ها در شیوه‌های گوناگون جست‌وجو در دو حوزه

نسبت‌ها	حوزه موضوعی	تعداد	میانگین	انحراف استاندارد	|t|	df	P
عبارتی دقیق ه کلیدواژه‌ای	پزشکی	387	90849.34	985444.11			
	لوم اجتماعی	682	5248.05	38419.35	2.266	1067	.024
عنوان دقیق به کلیدواژه	پزشکی	323	636036.74	1643141.78			
	لوم اجتماعی	579	1480592.04	5366069.39	-2.757	900	.006
نشانی دقیق به کلیدواژه	پزشکی	227	1367462.68	4174828.74			
	وم اجتماعی	414	4904793.23	22079173.32	-2.390	639	.017

همان‌گونه که مشاهده می‌شود، تفاوت معناداری از نظر نسبت بازیافت‌ها میان دو حوزه وجود دارد. مقدار p برای نسبت میانگین بازیافت‌های حاصل از جست‌وجوی "عبارتی دقیق به کلیدواژه"، جست‌وجوی "عنوان دقیق به کلیدواژه"، و جست‌وجوی "نشانی دقیق به کلیدواژه" به ترتیب عبارتست از ۰/۰۲۴، ۰/۰،۰۰۶، ۰/۰۱۷ که وجود تفاوت معنادار میان دو حوزه را تایید می‌کند. این یافته، به‌طور ضمنی، بیانگر آن است که بسط جست‌وجو با استفاده از واژه‌های غیرموضوعی (عمومی) و محدود کردن دامنه کاوش از جست‌وجوی کلیدواژه‌ای به سایر حالت‌ها نتایج دقیق‌تری را در حوزه پزشکی به دست می‌دهد. دلیل احتمالی این امر آن است که با وجود بسامد کمتر واژه‌های عمومی در حوزه پزشکی، استفاده از این واژه‌ها در عنوان و نشانی اینترنتی سایت‌های این حوزه در مقایسه با حوزه علوم اجتماعی بیشتر است. به‌عبارت دیگر، طراحان سایت‌های حوزه پزشکی، در مقایسه با حوزه علوم اجتماعی، واژه‌های عمومی را در عنوان و نشانی اینترنتی صفحات طراحی شده بیشتر به کار می‌برند. در عین‌حال، نتیجه کلی آنکه، برای بسط جست‌وجو در دو حالت یاد شده در این دو حوزه می‌توان از واژه‌های عمومی به‌خوبی استفاده کرد.

نتیجه‌گیری

زبان، مهم‌ترین محمل برای خلق، نشر، جست‌وجو و بازیابی محتوای منابع اطلاعاتی است. در عین‌حال، هر حوزه‌ای از دانش بشری زبان ویژه خود را دارد و از واژگان خاصی برای تولید دانش و برقراری ارتباط استفاده می‌کند. از این‌رو، پژوهش در زبان متون و تحلیل واژگان آنها همواره اهمیت فراوانی در بازنمون (نمایه‌سازی) و در نتیجه، در جست‌وجو و بازیابی اطلاعات دارد. در این پژوهش، تلاش شد تا رویکرد نوینی در بسط جست‌وجو، فراسوی شیوه‌های متعارف کاربرد اصطلاحنامه‌ها، مورد استفاده قرار گیرد. هدف مقاله، بررسی قابلیت‌های واژه‌ها و عبارت‌های عمومی برای بسط جست‌وجو و ارزیابی نتایج بازیابی از نظر میزان دقت بازیافت‌ها بود. این هدف از طریق تحلیل متن منابع اینترنتی در دو حوزه پزشکی و علوم اجتماعی، شناسایی واژه‌ها و عبارت‌های عمومی هر حوزه و نهایتا دسته‌بندی آنها از نظر میزان عمومی بودن و از نظر مکان (پیش و پس بودن) آنها

نسبت به واژه‌های موضوعی انجام گرفت. شناسایی واژه‌های عمومی که به‌طور معمول پیش یا پس از واژه‌های موضوعی در منابع اینترنتی ظاهر می‌شوند برای بسط جست‌وجو و بازیابی نتایج دقیق‌تر اهمیت دارد. یافته‌های پژوهش نیز نشان داد که با افزودن واژه‌های عمومی به‌عبارت جست‌وجو در موتور کاوش گوگل تعداد نتایج کمتر اما منابع دقیق‌تری بازیابی می‌شود. بر اساس این یافته، موتورهای کاوش می‌توانند سیاهه واژه‌های عمومی را برای بسط جست‌وجو در دسترس کاربران قرار دهند یا آنکه به‌صورت هوشمند عبارت جست‌وجوی کاربران را با واژه‌های عمومی ترکیب کرده و نتایج را برای استفاده احتمالی آنها نمایش دهند. با توجه به‌وجود تفاوت در واژگان عمومی حوزه‌های موضوعی گوناگون، موتورهای کاوش تخصصی می‌توانند واژه‌های عمومی با بسامد بالا را شناسایی کرده و سیاهه حاصل را در پایگاه خود تعبیه و در بسط جست‌وجو توسط کاربران مورد استفاده قرار دهند. با توجه به محدود بودن دامنه واژه‌های عمومی که همراه کلیدواژه‌های تخصصی در هر حوزه ظاهر می‌شوند، امکان شناسایی این واژه‌ها و تدوین یک سیاهه هوشمند از آنها به آسانی وجود دارد. طراحان صفحات وب را می‌توان تشویق نمود تا برای نامگذاری عنوان صفحات و نیز برای تدوین نشانی (URL) صفحات، از واژه‌ها و عبارت‌های عمومی و تخصصی معنادار (بامسمی)تری استفاده کنند تا میزان بازنمونی (نمایه‌پذیری) صفحات توسط موتورهای کاوش افزایش یابد. از سوی دیگر، طراحان موتورهای کاوش نیز می‌توانند در تدوین الگوریتم نمایه‌سازی صفحات و منابع اینترنتی، وزن بیشتری برای عنوان و نشانی اینترنتی صفحات وب در نظر گیرند. همچنین، آنها می‌توانند پیش‌فرض جست‌وجو را که به‌طور معمول بر جست‌وجوی کلیدواژه‌ای استوار است به‌عنوان و نشانی دقیق صفحات معطوف کنند تا نتایج دقیق‌تر و مرتبط‌تری حاصل شود.

پیشنهاد برای پژوهش‌های بیشتر در این زمینه

در زمینه شناسایی و دسته‌بندی واژه‌های عمومی که قابلیت استفاده در بسط جست‌وجو دارند می‌توان پژوهش‌های دیگری به انجام رساند:

۱. مشابه پژوهش حاضر، می‌توان پژوهشی با جامعه (تعداد صفحات) بسیار گسترده‌تری انجام داد تا تعداد و گستره بسیار بیشتری از واژه‌های عمومی را شناسایی و در بسط جست‌وجو مورد استفاده قرار داد. هدف چنین پژوهشی تایید یا رد یافته‌ها و نتایج پژوهش حاضر و نیز تکمیل سیاهه واژه‌های عمومی است.
۲. پژوهش‌های مشابهی را می‌توان در حوزه‌های موضوعی دیگر نیز انجام داد تا هم واژه‌های عمومی غیروابسته به حوزه‌های موضوعی (یعنی واژه‌های عمومی مشترک در همه حوزه‌ها) و هم واژه‌های عمومی موضوع–وابسته شناسایی شود.
۳. در زمینه شناسایی واژه‌های عمومی و دسته‌بندی آنها از نظر حالت دستوری (فعل، اسم، صفت، قید و غیره) می‌توان پژوهشی انجام داد تا از یافته‌های آن در جهت طراحی ابزارهای هوشمند برای کمک به بسط جست‌وجو به زبان طبیعی استفاده کرد.
۴. پژوهش دیگری می‌توان به‌صورت گسترده در زمینه میزان وجود واژه‌های عمومی در عنوان و نیز در نشانی اینترنتی صفحات وب انجام داد تا قابلیت این دو مورد را به‌منزله پیش‌فرض جست‌وجو در موتورهای کاوش و سنجش میزان بازیابی دقیق و مرتبط بررسی کرد.

نتایج این گونه پژوهش‌ها کمک خواهد کرد تا بتوانیم قابلیت‌های زبان طبیعی و به‌ویژه واژه‌های عمومی را در بسط جست‌وجو مورد سنجش دقیق قرار دهیم. بدیهی است تعبیه ابزارهای هوشمند در موتورهای کاوش مستلزم انجام پژوهش‌های بیشتر و به دست آوردن یافته‌های عینی‌تر است.

فهرست منابع

Anick, P. G. and Tipirneni, S. (1999). The paraphrase search assistant: terminological feedback for iterative information seeking, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, (pp.153-159), August 15-19, 1999, Berkeley, California, United States.

Baeza-Yates, R., Hurtado, C. and Mendoza, M. (2004). "Query Recommendation Using Query Logs in Search Engines" EDBT Workshops. (pp.588-596). Retrieved August 17, 2005 from http://www.dcc.uchile.cl/~churtado/clustwebLNCS.pdf

Billerbeck, B. and Zobel, J. (2003). "When query expansion fails" in: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM.

Bruza, P., McArthur, R. and Dennis, S. (2000). Interactive Internet search: keyword, directory and query reformulation mechanisms compared, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, p.280-287, July 24-28, 2000, Athens, Greece

Casasola, E and Gauch, S. (1997). Intelligent information agents for the World Wide Web. Technical Report ITTC-FY97-11100-1, Information and Telecommunication Technology Center, The University of Kansas.

Chan, Lois Mai, and others. (2001). "A faceted approach to subject data in the Dublin Core Metadata Record." Journal of Internet Cataloging, 4(1 / 2), 35-47.

Chowdhury, G. G. (1999). The Internet and information retrieval research: A brief review. Journal of Documentation, 55(2), 209-225.

Chowdhury, G.G. & Chowdhury, S. (1999). Digital library research: major issues and trends. Journal of Documentation, 55(4), 409-448.

Chowdhury, A., and Soboroff, I. (2002). Automatic Evaluation of World Wide Web Search Services. SIGIR'02, pp. 421-422. Retrieved October 15, 2005 from http://citeseer.ist.psu.edu/chowdhury02automatic.html

Dias, P., Gomes, M.J., and Correia, A.P. (1999). Disorientation in hypermedia environments: Mechanisms to support navigation. Journal of Educational Computing Research, 20(2), 93-117.

Doan, Khoa et al. (1997). Query Previews for Networked Information Systems: A Case Study with NASA Environmental Data. SIGMOD Record (ACM Special Interest Group on Management of Data).

Efthimiadis, E. N. (2000). Interactive query expansion: a user-based evaluation in a relevance feedback environment, Journal of the

American Society for Information Science, 51(11), 989-1003, Sept. 2000

Efthimiadis. E. N. (1995). User choices: a new yardstick for the evaluation of ranking algorithms for interactive query expansion, Information Processing and Management: an International Journal, 31(4), 605-620.

Ellis, D., Ford, N., and Furner, J. (1998). In search of the unknown user: Indexing and hypertext and the World Wide Vfeb. Journal of Documentation, 54(1), 28-47.

Filman and Pant, (1998). "Searching the Internet", IEEE Internet Computing, 2(4), 21-23.

Griffiths, J. and Brophy, P. (2005). "Student Searching Behavior and the Wcb: Use of Academic Resources and Google". Library Trends, 53(4), 539-578.

Harman, D. (1988).Towards interactive query expansion, Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval, pp.321-331, Grenoble, France

Henzinger, Motwani, and Silverstein. (2002). "Challenges in Web search engines", Colloquium Papers, Right Now Technologies, pp.1-12. Retrieved July 29, 2005 from http://ai.rightnow.com/colloquium/papers.php

Hicks, C., Rush, J., and Strong, S. (1977). Content analysis. In: Encyclopedia of Computer Science and Technology. New York: Jack Belzer, v.6.

Iivonen, M. (1995). Searchers and Searchers: Differences Between the Most and Least Consistent Searchers. In: Fox, E.A., Ingwersen, P. and Fidel, R. eds. Proceedings of the 18th annual international ACM SIGIR conference on research and development in information retrieval. New York, NY: ACM, 1995, 149-157.

Jansen, B. J., Spink, A. and Saracevic, T. (2000). "Real life, real users and real needs: A study and analysis of users queries on the Web." Information Processing and Management, 36(2), 207-227.

Lawrence & Giles. (1998). Context and Page Analysis for Improved Web Search, IEEE Internet Computing, 2(4), 38-46.

Lykke, M. and Ingwersen, P. (1999) The word association methodology - a gateway to work-task based retrieval. Retrieved September 17, 2005 from http://66.102.7.104/search?q=cache:d4V9LACbPeQJ:ewic.bcs.org/conferences/1999/mira99/papers/paper6.pdf+%22subject+searching%22+non-topical&hl=en

Lyons, J. Language and Linguistics: An Introduction. Cambridge University Press, 1981. (Reprint with new preface, 1995.)

McArthur, R. and Bruza, P.D. (2000). The Ranking of Query Refinements of Interactive Web-based Retrieval. Proceedings of the Information

Doors Workshop (held in conjunction with the ACM Hypertext and Digital Libraries Conferences).

Oyama, S. et al. (1999). Keyword Spices: A New Method for Building Domain-Specific Web search engines. Retrieved August 5, 2005 from http://www.lab7.kuis.kyoto-u.ac.jp/~ishida/pdf/ijcai01.pdf

Pokorny, J. (2004). "Web searching and information retrieval". IEEE Computer Software, 6(4), 43–48.

Soboroff, I. (2004). "On evaluating web search with very few relevant documents", Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. New York: ACM Press.

Spink, A., Greisdorf, H., and Bateman, J. (1998). From highly relevant to not relevant: Examining different regions of relevance. Information Processing and Management, 34(2/3), 257-274. Retrieved September 2, 2005 from http://www.informatik.uni-trier.de/ ~ley/db/indices/a-tree/s/Spink:Amanda.html

Spink, A. and Xu, J. L. (2000). "Selected results from a large study of Web searching: the Excite study", Information Research, 6 (1). Retrieved August 9, 2005 from http://informationr.net/ir/6-1/paper90.html

Sugiura, Atsushi and Etzioni, O. (2000). Query routing for Web search engines: architectures and experiments. In: Proceedings of the 9th international World Wide Web conference on Computer networks: the International Journal of Computer and Telecommunications Networking. Amsterdam: North-Holland: pp. 417-429.

Tillett, B. B. (2001). "Authority control on the Web" Retrieved August 2, 2005 from http://www.loc.gov/catdir/bibcontrol/tillett_paper.html

Voorbij, H. J. (1999). Searching scientific information on the Internet: A Dutch academic user survey. Journal of the American Society for Information Science, 50(7), 598-615.

پیوست یک: ۵۰ واژه/عبارت عمومی پربسامد در دو حوزه پزشکی و علوم اجتماعی

	Non-topical terms	NTT	STI	Location Before / Aft	Total Fr	Freq. in Sc Sci.	Freq. ir Health	Shared SS Health
1	about ~	X		B	78	40	38	X
2	~ is	X		A	72	47	25	X
3	~ and	X		A	63	44	17	X
4	~ prevention		X	A	49	11	38	X
5	and ~	X		B	40	15	25	X
6	national ~		X	B	40	22	18	X
7	~ information	X		A	25	9	14	X
8	information about ~	X		B	22	4	18	X
9	~ in	X		A	20	7	13	X
10	International ~		X	B	20	20	0	
11	~ research	X		A	15	2	13	X
12	~ issues	X		A	15	15	0	
13	Definition of ~	X		B	15	15	0	
14	Teen ~		X	B	15	0	15	
	~ foundation	X		A	14	12	2	X
16	~ reform		X	A	14	12	2	X
17	risk of ~		X	B	14	0	14	
18	~ of	X		A	13	10	3	X
19	~ cases		X	A	13	0	13	
20	Commission on ~		X	B	13	13	0	
21	Economic ~		X	B	13	13	0	
22	~ system		X	A	12	7	5	X
23	What is ~	X		B	12	10	2	X
24	~ products		X	A	12	0	12	
25	Against ~		X	B	11	10	1	X
26	Types of ~	X		B	11	10	1	X
27	~ awards		X	A	11	11	0	
28	diagnosed in ~		X	B	11	0	11	
29	Prescription ~		X	B	11	0	11	
30	To prevent ~		X	B	10	3	7	X
31	Nuclear ~		X	B	10	10	0	
32	~ statistics		X	A	9	7	2	X
33	Information on ~	X		B	9	3	6	X
34	~ program		X	A	9	9	0	
35	Diagnosed with ~		X	B	9	0	9	
36	The term ~	X		B	9	9	0	
37	~ articles	X		A	8	2	6	X
38	~ laws		X	A	8	6	2	X
39	~ news	X		A	8	6	2	X
40	~ survivors		X	A	8	1	7	X
41	Prevention of ~		X	B	8	6	2	X
42	Survivors of ~		X	B	8	2	6	X
43	~ attempt		X	A	7	0	7	

44	~ law		X	A	7	7	0
45	~ outbreak		X	A	7	0	7
46	~ patients		X	A	7	0	7
47	~ tips	X		A	7	0	7
48	Definitions of ~	X		B	7	7	0
49	Domestic ~		X	B	7	7	0
50	Radical ~		X	B	7	7	0

پیوست دو: نمونه‌ای از سیاهه واژه‌های عمومی که همراه کلیدواژه موضوعی "جهانی شدن" می‌آید

about **globalization**	**globalization** articles
Against **globalization**	**globalization** attempt
Articles on **globalization**	**globalization** awareness
Aspects of **globalization**	**globalization** basics
Basic information about **globalization**	**globalization** benefits
Benefits of **globalization**	**globalization** campaign
Books on **globalization**	**globalization** center
Combating **globalization**	**globalization** FAQs
Commission on **globalization**	**globalization** foundation
Counter **globalization**	**globalization** Guide
Definition of **globalization**	**globalization** information
Definitions of **globalization**	**globalization** information center
Economic **globalization**	**globalization** issues
Effects of **globalization**	**globalization** laws
FAQs about **globalization**	**globalization** links
History of **globalization**	**globalization** news
information about **globalization**	**globalization** principles
Information on **globalization**	**globalization** problems
International **globalization**	**globalization** process
national **globalization**	**globalization** programs
Prevention of **globalization**	**globalization** refers to
Radical **globalization**	**globalization** reform
Resources on **globalization**	**globalization** research
risk of **globalization**	**globalization** statistics
Survivors of **globalization**	**globalization** stories
To prevent **globalization**	**globalization** system
Types of **globalization**	**globalization** taxes
Understanding **globalization**	
Victim of **globalization**	
War on **globalization**	
What is **globalization**	