جستجو اطلاعات فارسي از اينترنت

خط فارسی دارای مشکلات مختلفی است که در جستجو و بازیابی اطلاعات، مسائل و مشکلات فراوانی را فراروی کابران اینترنت قرار می‌دهد. به خصوص با رشد سریع انتشارات الکترونیکی بر روی وب در شکل‌های مختلف پایگاه‌های اطلاعاتی، وبلاگ و... هیچ قاعده مشخص و ثابتی برای رسم‌الخط فارسی وجود ندارد و این مسأله باعث شده تا جستجوگران مطالب فارسی با مشکلات فراوانی روبرو شوند.

اینترنت به عنوان یک محمل اطلاعاتی عظیم، منابع اطلاعاتی را در مقیاسی وسیع در دسترس مخاطبان بالقوه قرار داده است. سهولت دسترسی به منابع اطلاعاتی اعم از متن و سایر رسانه‌ها عمده‌ترین مزیت اینترنت محسوب می‌شود.
این توانایی که هر کس ناشر آثار خود باشد عواقب ناخواسته‌ای را نیز در پی خواهد داشت و آشکارترین معضل، آن است که انبوهی از منابع بسیار متنوع و غیرقابل مدیریت را فراهم می‌آورد. افزایش سریع منابع اینترنتی نیازمند یک سازمان‌دهی مفید و مؤثر است. هرچند در حال حاضر راهنمای‌هایی برای منابع اینترنتی تهیه شده است که براساس فایل‌های مقلوب ساخته شدهِ توسط موتورهای جستجو و با استفاده از قابلیت‌های مختلف این موتورها از جمله؛ استفاده از عملگرهای بولی، جستجوی دقیق عبارت، محدود کردن یک جستجو به بخش خاصی از رکورد (مانند عنوان، آدرس) کوتاه‌سازی کلمات، جستجوی نزدیک‌یابی واژه‌ها، ایجاد محدودیت زمانی و منطقه‌ای و زبانی و .... به جستجوی اطلاعات کمک می‌کند اما باید تاکید کرد که در امر بازیابی اطلاعات از اینترنت بدون نمایه‌سازی نظام یافته نمی‌توان انتظار بازیابی مفید و مؤثر را داشت.
هرچند بیشترین اطلاعات موجود بر روی اینترنت به زبان انگلیسی است، ولی حجم اطلاعات به زبان فارسی نیز با سرعت در حال افزایش است و کاربران به دلایل مختلفی علاقه زیادی به اطلاعات فارسی نشان می‌دهند و از آن جایی که زبان غالب در اینترنت انگلیسی است، جستجو به زبان‌های غیرانگلیسی از جمله فارسی، مسایل و مشکلات مختلفی را جدای از مشکلات عمومی اینترنت دارد.

خط فارسی دارای مشکلات مختلفی است که در جستجو و بازیابی اطلاعات، مسائل و مشکلات فراوانی را فراروی کابران اینترنت قرار می‌دهد. به خصوص با رشد سریع انتشارات الکترونیکی بر روی وب در شکل‌های مختلف پایگاه‌های اطلاعاتی، وبلاگ و... هیچ قاعده مشخص و ثابتی برای رسم‌الخط فارسی وجود ندارد و این مسأله باعث شده تا جستجوگران مطالب فارسی با مشکلات فراوانی روبرو شوند.

اینترنت به عنوان یک محمل اطلاعاتی عظیم، منابع اطلاعاتی را در مقیاسی وسیع در دسترس مخاطبان بالقوه قرار داده است. سهولت دسترسی به منابع اطلاعاتی اعم از متن و سایر رسانه‌ها عمده‌ترین مزیت اینترنت محسوب می‌شود. این توانایی که هر کس ناشر آثار خود باشد عواقب ناخواسته‌ای را نیز در پی خواهد داشت و آشکارترین معضل، آن است که انبوهی از منابع بسیار متنوع و غیرقابل مدیریت را فراهم می‌آورد. افزایش سریع منابع اینترنتی نیازمند یک سازمان‌دهی مفید و مؤثر است. هرچند در حال حاضر راهنمای‌هایی برای منابع اینترنتی تهیه شده است که براساس فایل‌های مقلوب ساخته شدهِ توسط موتورهای جستجو و با استفاده از قابلیت‌های مختلف این موتورها از جمله؛ استفاده از عملگرهای بولی، جستجوی دقیق عبارت، محدود کردن یک جستجو به بخش خاصی از رکورد (مانند عنوان، آدرس) کوتاه‌سازی کلمات، جستجوی نزدیک‌یابی واژه‌ها، ایجاد محدودیت زمانی و منطقه‌ای و زبانی و .... به جستجوی اطلاعات کمک می‌کند اما باید تاکید کرد که در امر بازیابی اطلاعات از اینترنت بدون نمایه‌سازی نظام یافته نمی‌توان انتظار بازیابی مفید و مؤثر را داشت. هرچند بیشترین اطلاعات موجود بر روی اینترنت به زبان انگلیسی است، ولی حجم اطلاعات به زبان فارسی نیز با سرعت در حال افزایش است و کاربران به دلایل مختلفی علاقه زیادی به اطلاعات فارسی نشان می‌دهند و از آن جایی که زبان غالب در اینترنت انگلیسی است، جستجو به زبان‌های غیرانگلیسی از جمله فارسی، مسایل و مشکلات مختلفی را جدای از مشکلات عمومی اینترنت دارد.

حجم اطلاعات به زبان فارسی در روی اینترنت در اشکال مختلف آن به سرعت رشد کرده است. در حال حاضر توسعه وبلاگ‌های فارسی، سایت‌های علمی، تبلیغاتی و دانشگاهی به زبان فارسی باعث شده است که جایگاه زبان فارسی تا حد زبان اول ارتباطات اینترنتی نزد ایرانیان و فارسی‌زبانان در سراسر جهان ارتقا یابد. شاید بتوان گفت که اولین مرجع وبلاگ‌نویسی فارسی با انتشار راهنمای ساخت وبلاگ فارسی آغاز شده است. بدون شک دومین موج نیز با شروع به کار سایت پرشین بلاگ که امکان راه‌اندازی وبلاگ برای کاربران فارسی زبان را با سهولت بیش‌تری فراهم می‌کند آغاز شده است. اما پیامد قابل توجه دیگری که رشد وبلاگ‌نویسی در ایران داشته است پیدایش سایت‌های اینترنتی فارسی زبانی است که صاحبان وبلاگ‌ها ایجاد کرده‌اند و این خود موج جدیدی از گسترش کاربرد اینترنت در جامعه ایران به حساب می‌آید. اکنون روی آوردن برخی از روزنامه‌نگاران، پژوهش‌گران، دانشجویان به وب فارسی و استفاده از منابع خبری، علمی و .... موجب تقویت نقش رسانه‌ای وب فارسی شده است.

پدیده دیگری که باعث گسترش زبان و خط فارسی در اینترنت شده است، ایجاد کتابخانه‌های دیجیتالی فارسی در شبکه جهانی است. با این که از شکل‌گیری کتابخانه‌های فارسی در شبکه جهانی مدت زیادی نمی‌گذرد، اما با این حال به سرعت در حال رشد و گسترش است. شماری از این کتابخانه‌ها در پایگاه‌های اینترنتی شکل گرفته‌اند و بسیاری وبلاگ‌هایی هستند که برای این کار راه‌اندازی شده‌اند. از ویژگی‌های این کتابخانه‌ها این است که هیچ یک جنبه تجاری ندارند. آنچه در بسیاری از کتابخانه‌های مجازی فارسی در دسترس است تنها شامل کتاب نیست، بلکه نوشته‌هایی اعم از داستان، مقاله، تک نگاشت و .... نیز در میان مجموعه‌ها دیده می‌شود. هم چنین آثاری که احتمالاً هیچ گاه چاپ کاغذی ندارند و البته وجود کتاب‌هایی که مدت‌هاست نایاب هستند و مجال انتشار دوباره نیافته‌اند و یا آثاری که امروز به دلایلی بازچاپ آن‌ها مقدور نیست، از جاذبه‌های کتابخانه‌های مجازی‌اند. پایگاه اینترنتی کتاب‌های رایگان فارسی، پایگاه اینترنتی بانی تک، کتابخانه مجازی داستان‌های فارسی، آوای آزاد، پایگاه اینترنتی خوابگرد، کتابخانه دوات، پایگاه اینترنتی سخن، وبلاگ کتابخانه هرمس، پایگاه اینترنتی گفتمان، پایگاه تاریخ و فرهنگ ایران زمین، پایگاه مرکز جهانی اطلاع‌رسانی آل البیت، کتابخانه پایگاه اینترنتی حوزه، پایگاه اینترنتی امام علی (ع)، پایگاه اینترنتی کتابخانه دیجیتال و .... شماری از این کتابخانه‌ها هستند.

کاربران به دلایل مختلفی از قبیل «دسترسی آسان و ارزان به حجم عظیم اطلاعات، عدم نیاز اطلاعات یافته شده از اینترنت به تایپ مجدد، دسترسی سریع و اطلاعات جدید، صرفه‌جویی در وقت و عدم تسلط اکثر کاربران به زبان انگلیسی که زبان غالب بر اینترنت است» به دنبال اطلاعات فارسی از اینترنت هستند. گسترش زبان و انبوهی از نوشتارها ایجاب می‌کند که خط ضابطه داشته باشد و از سوی دیگر پیشرفت فن‌آوری و پیدایش اینترنت خواستار ضابطه و قانونمندی است. اطلاع‌رسانی که جنبه بین‌المللی پیدا کرده است بدون دستور خطی سامان یافته و نظام‌مند میسر نیست و دست‌کم بر دشواری‌ها می‌آفریند. در حال حاضر وبلاگ‌های فارسی مقام دوم یا سوم را در جهان دارا می‌باشد. به نظر دکتر آشوری، اگر زبان فارسی به همین صورت بی‌دقت در اینترنت به کار رود در سطح زبانی برای تفنن باقی خواهد ماند و کم‌تر حرفی جدی به این زبان زده خواهد شد. آینده زبان فارسی در اینترنت بستگی به این دارد که نویسندگان فارسی تا چه حد کار خود را جدی بگیرند و این زبان را بازسازی کنند که از لحاظ قدرت بیان و دقت مفاهیم و استواری ساختار دستوری به زبان انگلیسی نزدیک شود.

نبود استاندارد ثابت رسم‌الخط فارسی موجب این شده است که به تعداد صفحات وب فارسی سبک و سیاق نگارش به کار رفته باشد، لذا می‌توان چنین ارزیابی کرد که اکثر وب‌های فارسی در برخی خصوصیات مشترک می‌باشند از جمله این که نگارش برخی از آن‌ها زبان غیررسمی و محاوره‌ای است و به خصوص در متون علمی اغلب واژه‌های بیگانه به دفعات استفاده می‌شود. رسم‌الخط مورد استفاده نیز متفاوت و سلیقه‌ای است و برخی از آن‌ها غلط‌های تایپی و نگارشی فراوانی دارند و این خصوصیات، اغلب به جهت محدودیت‌های محیط الکترونیکی و عدم تطابق رسم‌الخط فارسی با آن می‌باشد که نمایه‌سازی و سپس جستجو به این زبان را با دشواری‌هایی رو به رو می‌سازد.

با توجه به این نکته که
اطلاعات ارزشمند فراوانی در اینترنت وجود دارد و اینترنت با شتابی فراوان به یک منبع اطلاعاتی ممتاز تبدیل شده است. موتورهای جستجو به عنوان یکی از اساسی‌ترین دروازه‌های ورود به منابع اینترنتی دارای ضعف‌هایی هستند که می‌توان به این موارد اشاره کرد:

- در یک مجموعه از یافته‌های بازیابی شده مدخل‌های تکراری فراوانی ملاحظه می‌شود.
- نتایج غیر قابل پیش‌بینی هستند.
-نتایج چه بسا گمراه کننده باشند؛ ممکن است جستجویی در یک موتور کاوش نتیجه‌ای نداشته، ولی در موتور دیگر دارای یافته‌های فراوان باشد.
-موتورهای کاوش محتویات پایگاه‌های اطلاعاتی خودشان را نشان نمی‌دهند و از معیارهایی که برای گنجاندن یک مدرک در فایل‌هایشان دارند حتی شرحی ارایه نمی‌کنند.
- مهار واژگانی وجود ندارد و قواعد نقطه‌گذاری و بزرگ‌نویسی نیز استاندارد نیست.
-
بدون بررسی عملی هر عنصر، اغلب نمی‌توان میزان ربط و رابطه‌ها را تحلیل کرد. یعنی اطلاعات کافی در مدخل نمایه نیست تا فرد بتواند دست به انتخاب بزند.
- عدم توان موتورهای جستجو در تمایز میان مدارکی که توسط فرد الف نوشته شده و مدارکی که درباره فرد الف نوشته شده است.
-منابع قابل توجهی در شبکه وب وجود دارند که توسط موتورهای جستجو نمایه نمی‌شوند. به این بخش از وب اصطلاحاً وب نامریی می‌گویند. «وب نامریی بخش بزرگی از وب است که موتورهای جستجو آن‌ها را نمی‌توانند نمایه کنند و عبارتند از: سایت‌های دارای رمز عبور، فایل‌های پی.دی.اف از متون آرشیو شده، ابزارهای تعاملی نظیر ماشین حساب‌ها و برخی از واژه‌نامه‌ها و

همچنین بعضی از پایگاه‌های اطلاعاتی، منابع محافظت شده از طریق اسم کاربر و گذر واژه، منابع و صفحات وب بدون پیوند و صفحات افزون بر حداکثر تعداد صفحات قابل مرور.»

جستجوی اطلاعات در اینترنت به دو روش می‌تواندصورت گیرد یکی استفاده از جملات زبان محاوره‌ای است و دیگری بکارگیریکلمات کلیدی. در روش استفاده از جملات زبان محاوره‌ای که اغلب به کاربران تازه‌کار پیشنهاد می‌شود. یکی از عیب‌های بزرگ این روش تعداد نتایج جستجوی زیادی است که بازگردانده می‌شود. به همین دلیل این روش توسط کاربران حرفه‌ای و حتی توسط همه، کمتر استفاده می‌شود.

یکی از کاراترین ومقتدرترین روش‌های جستجوی اطلاعات در دنیای وب استفاده از واژه‌هایی است که اصطلاحاً کلمات کلیدی نامیده می‌شوند. اغلب کاربران حرفه‌ای و جستجوگران ورزیده دنیای اینترنت می‌توانند با طرح بهترین کلمات کلیدی و بکار بستن قوانین ترکیب آن‌ها با هم برای نیازهای اطلاعاتی خود پاسخی در خور بیابند. در این روش توصیه‌های زیر برای انتخاب کلمات کلیدی و نیز جستجوی دقیق و مفید پیشنهاد می‌شود که بشرح ذیل است:

۱- حتی‌المقدور سعی شود کلمات کلیدی از میان اصطلاحات منحصر به فرد و اسامی خاص انتخاب شود.

۲- حتی‌المقدور از آوردن کلمات عمومی که عناوین بسیاری را در زیر مجموعه خود شامل می‌شوند، جداً خودداری کنید.

۳ – همیشه اسم شخص یا نام شی یا هر چیز دیگری را که مد نظر دارید به‌طور کامل وارد کنید.

۴ – دقت کنید که اگر موتور جستجو میان حروف بزرگ و کوچک تفاوتی می‌گذارد، این مسأله را در طرح کلمات کلیدی خود مدنظر داشته باشید.

۵ – در نظر داشته باشید اگر نتیجه جستجو صفر بود به احتمال زیاد می‌تواند از یک اشتباه تایپی باشد.

۶ – اگر املای صحیح و کامل کلمه‌ای را نمی‌دانید از کارکتر جانشین که اغلب * و یا ؟ است استفاده کنید.

7– اگر یک کلمه کلیدی را برای طرح دقیق و تمام و کمال یک مورد جستجو کفایت نمی‌کند، از تکنیک‌های جستجوی عبارتی، استفاده از اپراتورهای جبر بولین (AND, OR, NOT) استفاده کنید. جستجوی عبارتی یکی از مهمترین و قدرتمندترین امکانات جستجو در اغلب موتورهای جستجو می‌باشد و می‌توان یک

عبارت یا جمله مشخص را به همان ترتیبی که کلمات وارد شده‌اند مورد جستجو قرار داد. برای این روش جستجو عبارت مورد نظر را داخل گیومه"" بگذارید.

۸

- استفاده از عملگر AND : AND به مفهوم "و" برای محدود کردن دامنه جستجو از طریق ترکیب کلید واژه‌های مختلف به کار می‌رود و برای ترکیب کلیدهای جستجو زمانی که برای شما مهم است که دو یا چند کلمه کلیدی حتماً وجود داشته باشد و علامت آن در پایگاه‌های مختلف به صورت استفاده از عبارت AND، استفاده از + ، انتخاب عبارت ALL THE WORD از منو، انتخاب عبارت (ON MATCH ALL WORDS AND) به‌وسیله کلیک کردن بر روی دکمه‌های رادیویی است.

۹–

استفاده از عملگر OR: اپراتور OR به مفهوم "یا" و برخلاف عملگر AND باعث گسترش دامنه جستجو و بازیابی اطلاعات بیش‌تر شده برای ترکیب کلید واژه‌های جستجو زمانی که انتظار دارید تنها یک، دو یا چند کلمه کلیدی حضور داشته باشند و علامت آن استفاده از عبارت OR، نحوه‌ی اجرای ساده و معمولی آن، انتخاب عبارت ANY OF THE WORDS از منو، انتخاب عبارت (WORDS MATCH ON ANY OR) با کلیک بر روی دکمه‌های رادیویی می‌باشد. یکی از کاربردهای مهم این عملگر پوشش مفاهیم یا اصطلاحات مترادف، مرتبط یا با املاهای متفاوت است.

۱۰

– استفاده از عملگر NOT: اپراتور NOT به مفهوم "نه" و یا به جز که در این صورت تمامی جواب‌های بازگشتی که حاوی عبارت یا کلمه کلیدی هستند حذف خواهند گردید و برای اجرای آن تنها کافیست که NOT را قبل از عبارت یا کلمه کلیدی مورد نظرتان با یک فاصله بیاورید.

۱۱ – استفاده از

کوتاه‌سازی کلید واژه‌ها: این تکنیک به ما امکان می‌دهد که با وارد کردن بخشی از یک کلید واژه بتوانیم مشتقات مختلف آن را نیز در فرآیند جستجو بازیابی کنیم. اکثر موتورهای جستجو این تکنیک را با استفاده از علامت ستاره (*) ارایه می‌دهند. یکی از مشکلات استفاده از این تکنیک این است که باعث بازیابی اطلاعات غیرمرتبط و ناخواسته زیادی می‌شود.

– ۱۲

استفاده از عملگر نزدیک‌یابی: در بسیاری از موارد استفاده از عملگر AND باعث بازیابی اطلاعاتی می‌شود که برای ما مفید نیست. به این دلیل که این

عملگر کلید واژه‌ها را در هر کجای متن که باشند بازیابی می‌کند. در این موارد استفاده از تکنیک نزدیک‌یابی می‌تواند از ریزش کاذب اطلاعات و یا بازیابی اطلاعات غیرمرتبط جلوگیری نماید. همه موتورهای جستجو قابلیت استفاده از این تکنیک را ندارند ولی به عنوان مثال در موتور جستجوی آلتاویستا می‌توان با استفاده از عملگر NEAR از این تکنیک استفاده نمود.

۱۳
– جستجوی ترکیبی با استفاده از پرانتز: این تکنیک یکی از مهمترین تکنیک‌های جستجو می‌باشد که به وسیله آن می‌توان تا حدود زیادی از بازیابی موارد غیرمرتبط در محیط وب جلوگیری کرد. در این روش می‌توان از همه عملگرهای جستجو که در بالا گفته شده یک‌جا استفاده کرد و آن‌ها را با همدیگر ترکیب نمود.

۱۴ – جستجوی کلیدواژه در عنوان صفحات وب: این تکنیک با این پیش فرض که عنوان یک صفحه وب تا حدود زیادی نمایانگر محتوای اطلاعات موجود در آن است به جستجوی واژه‌های کلیدی در عنوان سایت‌ها می‌پردازد. علامت آن در موتورهای جستجو متفاوت است ولی اغلب موتورهای جستجو از طریق فهرست انتخابی و یا گزینه‌های دیگر این امکان را فراهم می‌آورند.

۱۵ – جستجوی حوزه سایت‌ها: با توجه به این که به صورت قراردادی هر کشوری حوزه خاصی در محیط وب دارد، قابلیت جستجوی حوزه سایت‌ها به ما این امکان را می‌دهد که فرایند جستجو را به حوزه خاصی نظیر سایت‌های وب ایران (IR) و یا سایت‌های وب سازمان‌های غیر انتفاعی (ORG) محدود کنیم. دستورات استفاده از این تکنیک در موتورهای جستجو مختلف می‌باشد.

۱۶ – محدود کردن جستجو به زبان‌های مختلف باعث می‌شود نتایج جستجو به زبان‌های دیگر آورده نشود و انتخاب مطلب مورد نظر آسان‌تر است.

۱۷
– محدود کردن جستجو به تاریخ انتشار منابع در وب: تاریخ انتشار یا به اصطلاح روزآمدی مطلب به خصوص در منابع علمی اصل مهمی است و این‌گونه محدودیت باعث می‌شود بنا به نیاز کاربر جدیدترین و یا قدیمی‌ترین منبع بازیابی بشود.

۱۸ – جستجوی رسانه‌های مختلف؛ موسیقی، عکس، ویدئو: زمانی که فقط نوع خاصی از رسانه مورد نیاز است به عنوان مثال زمانی که به

عکس یک شخصیت نیاز داریم، جستجو در میان عکس‌ها باعث می‌شود نتیجه جستجو شامل اطلاعات دیگری در مورد آن شخصیت نباشد.

۱۹ – جستجوی صفحات با
فرمت‌های مختلف: PDF, WORD, MP3, MPEG, ...: زمانی که فرمت خاصی مورد نظر است می‌توان از این تکنیک استفاده کرد . به عنوان مثال اگر مایل باشیم منبع بازیابی شده در فرمت PDF باشد، این تکنیک می‌تواند مفید باشد.

۲۰
– آگاهی از پیش‌فرض‌های جستجو در موتور جستجو: با توجه به این که هر موتور جستجو برای ترکیب واژه‌ها یک پیش‌فرض دارد و اگر از هیچ گونه عملگری استفاده نشود، کلید واژه‌ها را به صورت پیش‌فرض با یکی از عملگرهای جبر بولی ترکیب می‌کند؛ آگاهی از این پیش‌فرض موتورهای جستجوی مختلف مهارت ما را در جستجو بالا می‌برد.

۲۱ – وب نامریی: وب نامرئی به دو دلیل
کمی و کیفی اهمیت دارد کمی از این نظر که موتورهای جستجو فقط قادر هستند حدود ۱۶ درصد از اطلاعات موجود در اینترنت را بازیابی کنند و اندازه وب نامریی تقریبا ۵۰۰ برابر وب مریی است و کیفی از این نظر که منابع اطلاعاتی موجود در وب عمیق معمولا ارزشمند و مفید هستند و در بسیاری از موارد پاسخ‌گوی نیاز کاربران می‌باشند. آشنایی با ابزارهایی که برای شناسایی منابع وب نامریی به وجود آمده‌اند و کاربران را به سایت‌های مناسب راهنمایی می‌کنند، باعث دسترسی به این بخش عظیم از اطلاعات مفید و ارزشمند می‌شود. مثل سایت INVISIBLEWEB که فهرستی از منابع نامریی را و سایت COMPLETEPLASET که فهرستی از تقریبا ۴۰۰۰۰ پایگاه اطلاعاتی وب نامریی را ارایه می‌دهد.