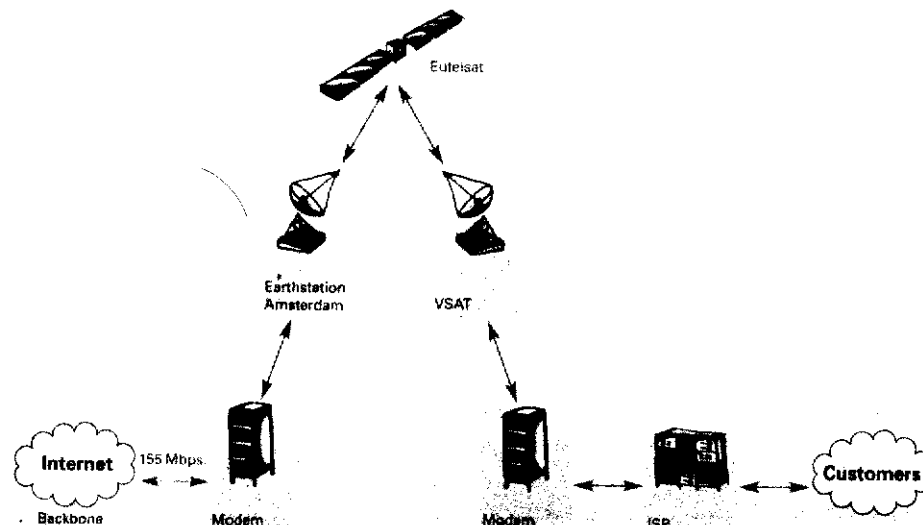




ابزارهای جستجو در اینترنت



نوشته : فرشته داودیان

کارشناس ارشد کتابداری و اطلاع رسانی

مقدمه

شکل چاپی نمی‌باشد. انتشار الکترونیکی بیش از ۹۳ درصد اطلاعات تولید شده در سال ۱۹۹۹ نمایانگر رواج اطلاعات در قالب‌های الکترونیکی است. (۱)

در حال حاضر شبکه جهان‌گستر اینترنت مناسبترین و پرکاربردترین بستر انتقال اطلاعات الکترونیکی در سراسر دنیا است. اینترنت حاوی میلیون‌ها فایل اطلاعاتی به اشکال گوناگون متنی، عددی، تصویری و صوتی است. تعداد این فایل‌ها به سرعت در حال افزایش است و با گذشت هر سال تقریباً دو برابر می‌شود. (۱) از طریق اینترنت می‌توان به اطلاعات بسیار متنوعی از قبیل متن کامل مقالات علمی، برنامه‌های آموزشی دانشگاه‌ها، برنامه گردهمایی‌های علمی، برنامه‌های شبکه‌های مختلف رادیویی و تلویزیونی، فهرست مندرجات و متن کامل بسیاری از مقالات، گزارش‌های علمی و حتی کتابها دست یافت.

از مهمترین پدیده‌های عصر حاضر نشر الکترونیکی بر بستر شبکه جهانی اینترنت می‌باشد. متداول‌ترین ابزارهای بازیابی اطلاعات بر این بستر فهرست‌های موضوعی، موتورهای جستجو و ابرموتورهای جستجو می‌باشند که هر یک برای گزینش، سازماندهی و بازیابی منابع اطلاعاتی مورد درخواست کاربران از تکنیک‌های خاص خود بهره گرفته‌اند. تشریح اندیشه حاکم بر مکانیسم عمل این ابزارهای، کاربران را با شیوه‌های بهینه بهره‌گیری از آنها آشنا می‌سازد.

امروزه یکی از مهمترین و متداول‌ترین روش‌های انتشار اطلاعات، انتشار اطلاعات به شکل الکترونیکی یا رقومی است. حجم بسیار زیاد اطلاعات منتشر شده به شکل الکترونیکی در سالهای اخیر به هیچ وجه قابل مقایسه با حجم اطلاعات منتشر شده به



اطلاعات قابل دسترس از طریق اینترنت در قالب صفحات وب^۱ منتشر می‌شوند. برای دسترسی به این اطلاعات، ابتدا می‌بایست به نحوی صفحه محل استقرار این اطلاعات شناسایی شوند. محل فیزیکی این صفحات بر روی کامپیوترهایی است که سرور^۲ یا سرورس دهنده نامیده می‌شوند و در نقاط مختلفی از سراسر جهان پراکنده‌اند (۱). به همین دلیل کامپیوتر ما نمی‌تواند برای انجام هر جستجو با مراجعه مستقیم به اصل صفحات، اطلاعات مورد نظر را بازیابی کند. با این شرایط چگونه می‌توان در این شبکه جهان‌گستر به سوی اطلاعات مورد نیاز و صفحات وب مورد نظر رهنمون شد؟

برای حل این مشکل ابزارهای جستجو در اینترنت بوجود آمده‌اند. این ابزارها که پایگاه‌های اطلاعاتی واسطه^۳ و یا سرورس‌های جستجو^۴ نامیده می‌شوند، در واقع مجموعه‌ای از اطلاعات منظم شده صفحات وب را در برمی‌گیرند (۸). ما نیز هنگام جستجو در اینترنت در واقع مستقیماً به یک یک سایت‌ها^۵ و صفحات وب مراجعه نمی‌کنیم، بلکه تنها در بین فهرست‌ها و نمایه‌های مختلف این پایگاه‌های اطلاعاتی به جستجو می‌پردازیم. بسیاری از ابزارهای جستجوی جهانی موجود در اینترنت آمریکایی هستند، همچنین منشأ اولیه و مبدأ اغلب سرورس‌های جستجوی جهانی در سراسر دنیا نیز سایت‌های آمریکایی می‌باشند. در اوایل سال ۱۹۹۷، در حدود ۶۰۰ سرورس جستجو در اینترنت وجود داشت. این تعداد در ابتدای سال ۲۰۰۰ میلادی به بیش از ۲۵۰۰ سرورس جستجوی فعال رسیده بود. (۱) سرورس‌های جستجو به چند گروه عمده فهرست‌های موضوعی^۶، موتورهای جستجو^۷، ابرموتورهای جستجو^۸ و دروازه‌های موضوعی^۹ تقسیم می‌شوند. اغلب افراد اصطلاح موتور جستجو را بطور عام برای همه انواع

ابزارهای جستجو در اینترنت به کار می‌برند، در صورتیکه این ابزارها کاملاً یکسان نیستند و تفاوت‌های بارزی با یکدیگر دارند. در این مقاله تنها فهرست‌های موضوعی، موتورهای جستجو و ابرموتورهای جستجو بررسی می‌شوند و معرفی ویژگی‌های دروازه‌های موضوعی که ابزارهای بسیار مفید و کارآمدی برای بازیابی اطلاعات تخصصی در زمینه‌های متنوع موضوعی می‌باشند، در فرصتی دیگر انجام خواهد شد.

فهرست‌های موضوعی:

هر فهرست موضوعی مجموعه‌ای از سایت‌ها و صفحات وب است که توسط افراد متخصص و متبحر آموزش دیده گردآوری و سازماندهی شده‌اند. در این فهرست‌ها سایت‌های اینترنتی در یک نظام سلسله مراتبی موضوعی گروه‌بندی می‌شوند. برای سازماندهی سلسله مراتبی اطلاعات ابتدا چند گروه موضوعی کلی و اصلی تعیین می‌شوند. این موضوعات مرحله به مرحله با تشکیل شاخه‌های فرعی، خاص‌تر می‌شوند. این نوع تقسیم‌بندی به موضوعات فهرست‌های موضوعی ساختاری هر می‌دهد. به دلیل همین نوع ساختار سلسله مراتبی، فهرست‌های موضوعی درخت‌های موضوعی^{۱۰} نیز نامیده می‌شوند.

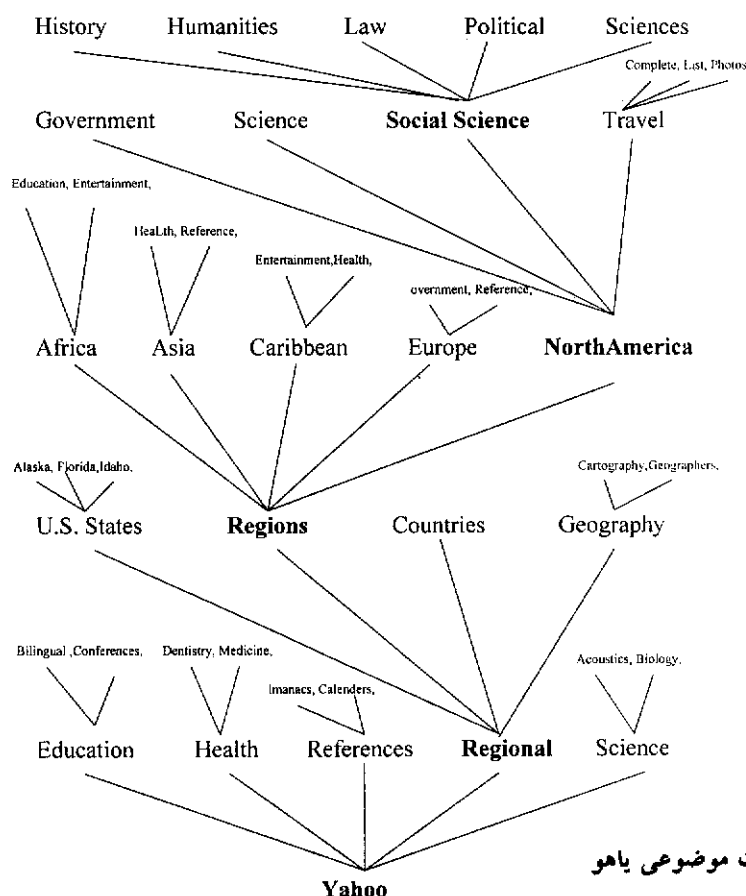
هر فهرست موضوعی برای دسته‌بندی صفحات وب از گروه‌بندی خاص خود استفاده می‌کند (۱). به همین دلیل گروه‌های موضوعی در همه فهرست‌های موضوعی یکسان نمی‌باشند. یاهو^{۱۱} که معروفترین و پراستفاده‌ترین فهرست موضوعی در اینترنت است، در حال حاضر دارای بیش از ۲۰۰۰ گروه موضوعی می‌باشد. زمانی که تعداد سایت‌های فهرست شده در یک گروه موضوعی یاهو به حدود ۱۰۰۰ سایت می‌رسد، گروهی از متخصصین موضوعی یک یا دو زیرشاخه برای آن موضوع در نظر می‌گیرند. نمودار



شماره یک (صفحه ۴) بخشی از تقسیمات موضوعی یاهو را نمایش می‌دهد.

ثبت اطلاعات سایت‌های اینترنتی در فهرست‌های موضوعی بطور کاملاً خودکار و ماشینی انجام نمی‌گیرد بلکه گروهی متشکل از افراد متخصص اغلب سایت‌هایی را که مرکز معرفی سایتها و صفحات جدید وب هستند، بطور مداوم تحت نظر دارند و سایت‌های جدید مناسب برای ثبت در فهرست موضوعی را به تشخیص خود انتخاب می‌کنند. طبیعی است که با بکارگیری چنین روشی احتمال ثبت سایت‌های قوی با مندرجات اطلاعاتی مفید و مناسب در یک فهرست موضوعی به مراتب بیش از سایت‌های ضعیف اطلاعاتی است. به این ترتیب اطلاعات کاملاً گزینشی و با نظارت افراد متخصص به مجموعه پایگاه

اطلاعاتی فهرست موضوعی افزوده و یا از آن حذف می‌شود. معمولاً پس از بررسی، متن بسیار کوتاهی برای معرفی سایت‌های منتخب تهیه و در پایگاه اطلاعاتی فهرست موضوعی، ذیل آدرس سایت ارائه می‌شود. موارد جستجوی استفاده‌کنندگان از این فهرست‌های موضوعی تنها در بین واژه‌های متن این توضیحات و معرفی‌های مندرج در پایگاه اطلاعاتی فهرست موضوعی، جستجو می‌شود. به دلیل هزینه و زمان زیاد مورد نیاز برای نمایه‌سازی صفحات وب توسط نیروی انسانی متخصص، معمولاً فهرست‌های موضوعی فاقد پوشش جامعی از منابع اطلاعاتی و مدارک اینترنتی می‌باشند، به همین دلیل در مقایسه با موتورهای جستجو معمولاً مدارک کمتری را تحت پوشش دارند.



نمودار شماره (۱): تقسیمات موضوعی یاهو



لازم به ذکر است که هر فردی می‌تواند سایت و صفحه وب خود را به یک فهرست موضوعی معرفی کند، ولی این امکان به هیچوجه به این معنا نیست که مشخصات سایت و صفحات وب معرفی شده حتما در فهرست موضوعی مزبور ثبت خواهد شد، بلکه این کار تنها با حذف مدت زمان مورد نیاز برای شناسایی سایت توسط کارشناسان فهرست موضوعی، موجب سرعت بخشیدن به بررسی سایت و اظهار نظر کارشناسان مربوطه در خصوص ثبت و یا عدم ثبت سایت مورد نظر در فهرست موضوعی می‌شود.

فهرست‌های موضوعی می‌توانند آغاز مناسبی برای آشنایی با اینترنت، و سایت‌هایی کارآمد برای جستجوی اطلاعات توسط کاربران تازه‌کار اینترنت باشند. این فهرست‌ها برای جستجوهای عمومی بسیار مناسبند. در صورتی که اطلاعات مورد نیاز ما بطور مشخص در یکی از گروه‌های موضوعی این فهرست‌ها قرار گیرد، استفاده از فهرست‌های موضوعی بسیار ساده و مفید خواهد بود. برای مثال زمانی که در جستجوی اطلاعاتی عمومی و کلی در مورد مؤسسات آموزشی باشیم، به سادگی می‌توانیم اطلاعات مورد نیاز خود را با استفاده از یک فهرست موضوعی بدست آوریم.

بر عکس هنگامی که موضوع مورد نظر ما بطور مشخص در یکی از تقسیمات موضوعی قرار نگرفته باشد، بازیابی اطلاعات مورد نظر از طریق فهرست‌های موضوعی با دشواری همراه خواهد بود.

فهرست‌های موضوعی در مقایسه با موتورهای جستجو منابع و مدارک اطلاعاتی کمتری را در خصوص موضوعات مورد جستجو بازیابی می‌کنند. در عوض مدارک بازیابی شده از طریق فهرست‌های موضوعی در مقایسه با منابع بازیابی شده توسط موتورهای جستجو، معمولا ارتباط بیشتری با موضوع

مورد جستجو دارند.

برخی از ویژگی‌های فهرست‌های موضوعی به اختصار عبارتند از:
- گزینش مدارک تحت پوشش فهرست‌های موضوعی با نظارت مستقیم نیروی انسانی متخصص انجام می‌گیرد.

- مدارک ثبت شده، در گروه‌های مشخص موضوعی و با ساختاری سلسله مراتبی سازماندهی می‌شوند.

- از آنجا که صفحات وب بصورت تمام متن در فهرست‌های موضوعی نمایه نمی‌شوند، تنها توصیفگرهای معین شده برای هر صفحه وب (مدرک) و یا واژه‌های مندرج در متن خلاصه توصیفی مدرک اینترنتی قابل جستجو و بازیابی می‌باشند.

اسامی و آدرس‌های اینترنتی^{۱۳} برخی از فهرست‌های موضوعی عبارتند از:

About.com	http://www.about.com
Academic Info	http://www.academicinfo.net
Argus Clearing House	http://www.clearinghouse.com
Ask Jeeves	http://www.askjeeves.com
Beyond the	http://www.blackstump.com.au
EdNA	http://www.edna.edu.au/EdNA
Education World	http://www.education-world.com
Librarian's Index	http://www.lii.org
Infomine	http://www.infomine.ucr.edu
Matilda	http://www.aamatilda.com

موتورهای جستجو

موتورهای جستجو سرویس‌دهندگان اینترنتی^{۱۴}

هستند که با گردآوری اطلاعات مدارک و منابع اینترنتی، امکان جستجو و بازیابی این مدارک و منابع اطلاعاتی را فراهم می‌آورند. موتورهای جستجو



اطلاعات مورد نیاز خود را با استفاده از نرم افزارهای ویژه، و با کمترین میزان دخالت مستقیم نیروی انسانی گردآوری و منظم می کنند.

ساختار موتورهای جستجو از نظر اندازه، امکانات جستجو، سیاست های نمایه سازی دقت و شکل ظاهری صفحات جستجو با یکدیگر متفاوتند. اغلب موتورهای جستجو اطلاعات میلیونها مدرک اینترنتی را گرد آورده اند، با وجود این حتی بهترین و عظیم ترین موتورهای جستجو نیز نمی توانند بیش از نیمی از مدارک موجود در اینترنت را پوشش دهند.

هر موتور جستجو دارای سه جزء اصلی می باشد: نرم افزار خزنده و جستجوگر، مجموعه و نرم افزار جستجو.

نرم افزار خزنده و جستجوگر:

این نرم افزار کارتک^{۱۴}، خزنده^{۱۵} و یا ربات^{۱۶} نامیده می شود. کارتکها با خزیدن بین صفحات وب، ضمن مرور این صفحات یک کپی از صفحات جدید را به پایگاه اطلاعاتی موتور جستجو اضافه می کنند. کارتکها همچنین با تعقیب پیوندهای هر یک از صفحات هر سایت می توانند تمام صفحات دیگر همان سایت را شناسایی و برای موتور جستجو ثبت کنند. از طرف دیگر با تعقیب پیوندها، کارتکها ممکن است سایت های جدید دیگری را نیز شناسایی نمایند. (۱)

کارتکها بطور مداوم و مستمر در دوره های زمانی مشخص، مثلاً هر ماه و یا هر دو ماه یکبار صفحات وب شناسایی شده را مجدداً بررسی می کنند. این بررسی های مکرر به منظور شناسایی صفحات ویرایش شده و ثبت تغییرات این صفحات در پایگاه اطلاعاتی موتور جستجو انجام می گیرند. چنانچه تغییری در صفحات وب ایجاد شود، موتور جستجو با

شناسایی و ثبت این تغییرات توسط کارتکها پایگاه اطلاعاتی خود را روزآمد می نماید. البته انعکاس این تغییرات همیشه بلافاصله پس از ویرایش صفحات وب در ایندکس های موتور جستجو ثبت نمی شوند. بدیهی است که در فاصله زمانی بین شناسایی یک صفحه وب (و یا تغییرات صفحه) و ثبت آن در موتور جستجو، بازبایی این صفحات (و یا تغییرات آنها) توسط موتور جستجوی مزبور امکان پذیر نیست. ثبت اطلاعات هر سایت اینترنتی در موتورهای جستجو نیز نظیر فهرست های موضوعی، به دو طریق می تواند انجام گیرد:

۱- معرفی سایت به موتور جستجو توسط پدیدآورندگان سایت. در صفحه اصلی^{۱۷} اغلب موتورهای جستجو گزینه ای وجود دارد که با انتخاب آن امکان معرفی صفحات وب و یا سایت های جدید به موتور جستجو وجود دارد. پدیدآورندگان صفحات و سایت های وب می توانند اطلاعات سایت خود را به پایگاه های اطلاعاتی موتور جستجو اضافه کنند. به این ترتیب اطلاعات یک سایت وب می تواند به محض انتشار در اینترنت در موتور جستجو ثبت شود.

۲- شناسایی سایت توسط نرم افزار خزنده موتور جستجو. در صورتیکه صفحه و یا سایت وب توسط تهیه کننده صفحه به موتور جستجو معرفی نشود و یا پیوندی از صفحاتی که قبلاً در موتور جستجو ثبت شده اند به این صفحه جدید وجود نداشته باشد، موتور جستجو بوسیله کارتکها این صفحات جدید را شناسایی و به مجموعه خود می افزاید.

مجموعه:

صفحات ثبت شده در موتور جستجو در کنار هم مجموعه موتور جستجو را تشکیل می دهند. این بخش که گاهی فهرست^{۱۸} نیز نامیده می شود، در اصل



یک کتاب غول‌پیکر است که حاوی یک نسخه از هر صفحه وب شناسایی شده توسط کارت‌نک می‌باشد. با ورود هر صفحه وب به مجموعه موتور جستجو، اطلاعات این صفحه بطور خودکار نمایه‌سازی شده و مدارک گردآوری شده از نظر موضوعی طبقه‌بندی می‌شوند. به این ترتیب مدارک گردآوری شده توسط موتور جستجو سازماندهی می‌شوند.

موتورهای جستجوی مختلف از روش‌های گوناگونی برای نمایه‌سازی مدارک اطلاعاتی استفاده می‌کنند. یکی از شیوه‌های مورد استفاده موتورهای جستجو برای نمایه‌سازی مدارک گزینش کلیدواژه‌ها با استفاده از متاتگها^{۱۹} است. زبان (HTML) این امکان را فراهم ساخته‌است که طراحان و تهیه‌کنندگان صفحات وب واژه‌های مورد نظر خود را به شیوه خاصی علامتگذاری و مشخص نمایند. این کلیدواژه‌های علامتگذاری شده اصطلاحاً متاتگ نامیده می‌شوند. آنجا که در حال حاضر بعضی از موتورهای جستجو مدارک اینترنتی را بر اساس متاتگها نمایه می‌کنند، تهیه‌کنندگان صفحه وب می‌توانند تأثیر مستقیمی بر نحوه نمایه شدن صفحه وب خود در این دسته از موتورهای جستجو داشته باشند. چنین امکانی مزایا و معایبی را نیز در پی دارد. طبیعی است که هیچ کس به اندازه پدیدآورنده یک صفحه وب نمی‌تواند از محتوای آن صفحه اطلاع داشته باشد. بنابراین در صورت انتخاب کلیدواژه‌هایی که بیانگر محتوای واقعی صفحه وب هستند و تبدیل این کلید واژه‌ها به متاتگ گام مهمی در جهت افزایش دقت نمایه‌سازی مدارک اینترنتی برداشته خواهد شد. از طرف دیگر اغلب بنا به دلایل تجاری و به منظور جلب افراد به برخی از صفحات وب، تهیه‌کنندگان این صفحات ممکن است کلماتی را در صفحه وب به متاتگ تبدیل کنند که با وجود جذاب بودن بیانگر محتوای واقعی صفحه و

سایت وب نباشند. در چنین صورتی متاتگها به جای کمک به شناسایی دقیق‌تر محتوای مدارک اینترنتی مانند یک تله عمل کرده و موجب جلب افراد به سایت‌هایی نامرتبط با موضوع مورد جستجوی آنها شوند.

میزان توجه به متاتگها و استفاده از این عناصر برای نمایه‌سازی صفحات وب در همه موتورهای جستجو یکسان نمی‌باشد. برخی از موتورهای جستجو تأکید بسیاری بر استفاده از متاتگها در نمایه‌سازی مدرک اینترنتی دارند، در صورتیکه بعضی دیگر از موتورهای جستجو برای این منظور کوچکترین استفاده‌ای از متاتگها نمی‌کنند. برای مثال HotBot و Infoseek اهمیت کمی به متاتگها می‌دهند، Lycos نیز در سازماندهی اطلاعات پایگاه اطلاعاتی خود هیچ توجهی به متاتگها نمی‌کند.

در صورتی که تهیه‌کنندگان سایت‌های وب با استفاده از متاتگها در نمایه‌سازی مدارک اینترنتی خود دخالت نکنند، موتورهای جستجو خود با بکارگیری نرم‌افزارهای ویژه‌ای واژه‌های کلیدی هر مدرک را تعیین و انتخاب می‌کنند. عوامل مختلفی می‌توانند میزان اهمیت یک واژه از نظر نرم‌افزارهای نمایه‌ساز موتورهای جستجو را تعیین نمایند. از جمله این عوامل محل قرار گرفتن واژه در مدرک و دفعات تکرار واژه است.

محل قرار گرفتن کلمات نیز از عواملی است که می‌تواند بیانگر اهمیت واژه باشد. کلیدواژه‌هایی که در عنوان، متاتگها و یا فیلدهای توصیفی مدرک حضور دارند، در مقایسه با کلیدواژه‌هایی که بطور معمولی در بخش‌های دیگر متن مدرک وجود دارند از اهمیت بیشتری برخوردارند و معمولاً به عنوان کلیدواژه‌های مهم نمایه می‌شوند. استدلال استفاده‌کنندگان از چنین شیوه‌ای در نمایه‌سازی مدارک این است که حضور کلیدواژه در ابتدای متن مدرک، عنوان و یا سر فصلهای



مدرک، بیانگر ارتباط موضوعی مدرک با کلیدواژه‌های مورد جستجو می‌باشد (۱). برای مثال موتور جستجوی Lycos بر اساس دفعات تکرار کلیدواژه مورد نظر شما در مدرک، همچنین محل قرار گرفتن کلیدواژه‌ها از قبیل عنوان سرفصلها و یا متن مدرک، کلیدواژه‌های مهم مدرک را شناسایی می‌کند. از طرف دیگر بعضی از موتورهای جستجو کلیدواژه‌های مهم مدارک را تنها از بخشهای ویژه‌ای از مدرک استخراج می‌کنند. برای مثال Lycos تنها کلیدواژه‌های عنوان مدرک، سرفصلها و همچنین بیست پیوند^۲ اولیه هر سایت به دیگر سایتها را نمایه می‌کند. در صورتی که Infoseek از روش نمایه‌سازی تمام متن استفاده می‌کند، در این روش کلیه کلمات متن مدرک، بدون توجه به محل قرار گرفتن و وزن آنها نمایه می‌شوند. البته در اینگونه موارد اغلب موتورهای جستجو مثل Infoseek و HotBot کلماتی مثل a, an, the, is, www که topords نامیده می‌شوند را نمایه نمی‌کنند.

دفعات تکرار واژه در متن و یا وزن واژه نیز می‌تواند عامل دیگری در تعیین اهمیت واژه‌های متن مدرک از نظر برخی از موتورهای جستجو باشد. در چنین حالتی واژه‌هایی که چند بار در طول متن مدرک تکرار شده‌اند مهمتر فرض شده و نمایه می‌شوند. بسامد بالای واژه در متن همواره نمایانگر اهمیت واژه نمی‌باشد، بلکه در مواردی نیز تکرار بیش از حد واژه‌ها در یک صفحه وب باعث می‌شود که موتور جستجو به آن واژه توجهی نکرده، واژه را به ایندکس خود اضافه نکند.

پیوندهای منتهی به مدرک نیز از نکاتی است که می‌تواند بیانگر درجه اهمیت مدرک باشد. برخی از موتورهای جستجو برای طبقه‌بندی موضوعی و تعیین اهمیت صفحات وب از شیوه‌های دیگری نیز استفاده می‌کنند. برای مثال Lycos و Excite تعداد پیوندهای

منتهی به مدرک را محاسبه و به نسبت تعداد آنها امتیازی برای مدرک قائل می‌شوند، هر چه تعداد این پیوندها بیشتر باشد درجه اهمیت صفحه نیز بیشتر خواهد بود. استفاده‌کنندگان از چنین شیوه‌ای معتقدند این امر که مدرکی چنان مورد توجه دیگران قرار گیرد که از صفحات وب خود پیوندهای منتهی به آن ایجاد کنند، می‌تواند نشانگر اهمیت و اعتبار مدرک مورد نظر باشد.

علاوه بر موارد فوق موتورهای جستجو ممکن است تفاوت‌های دیگری نیز در نمایه‌سازی مدارک اینترنتی با یکدیگر داشته باشند. برای مثال برخی از موتورهای جستجو بین حروف کوچک و بزرگ تفاوتی قائل نمی‌شوند و به هنگام نمایه‌سازی واژه‌های یکسانی که با حروف بزرگ یا کوچک تاپ شده باشند را معادل فرض می‌کنند. بعضی دیگر از موتورهای جستجو نیز بین حروف کوچک و بزرگ تفاوت قائل می‌شوند.

با توجه به موارد فوق طبیعی است که موتورهای جستجوی مختلف دارای مجموعه‌های مشابه نباشند، زیرا:

اولاً تعداد صفحات و مدارک ثبت شده در هر موتور با موتورهای جستجوی دیگر متفاوت است.
ثانیاً سیاست‌های نمایه‌سازی مدارک در همه موتورهای جستجو یکسان نیست.

نرم‌افزار جستجو:

نرم‌افزار جستجو بخش سوم هر موتور جستجو را تشکیل می‌دهد. این نرم‌افزار یک برنامه کامپیوتری است که با بررسی دقیق میلیونها صفحه وب ثبت شده در مجموعه موتور جستجو، نتایج مناسب هر جستجو را برگزیده و بر اساس میزان ارتباط با سؤال طرح شده درجه‌بندی می‌کند.



هنگامی که یک جستجوگر سؤالی از موتور جستجو می‌پرسد، موتور جستجو در واقع پایگاه اطلاعاتی از پیش ساخته شده خود را جستجو می‌کند. به عبارت دیگر بر خلاف تصور اکثر افراد، موتورهای جستجو برای بازیابی اطلاعات تمام سایت‌های موجود در اینترنت را جستجو نمی‌کنند، بلکه تنها در بین اطلاعات از قبل ثبت شده در پایگاه اطلاعاتی خود منابع و مدارک مرتبط را بازیابی می‌کنند. در هر جستجو با توجه به اجزای سؤال طرح شده، منابع موجود در پایگاه اطلاعاتی موتور جستجو از لحاظ میزان ارتباط موضوعی با سؤال مطرح شده، درجه‌بندی می‌شوند. سپس فهرستی از صفحات و منابعی که بر اساس این درجه‌بندی دارای کلیه شرایط مورد نظر جستجوگر باشند، ارائه می‌شوند (۱). مدارکی که بیشترین درجه ارتباط را با مورد جستجو دارا باشند در صدر فهرست منابع بازیابی شده آورده می‌شوند. در برخی از موتورهای جستجو می‌توان شکل و نحوه نمایش نتایج یک جستجو را نیز انتخاب و مشخص نمود.

از آنجا که این پایگاهها اغلب بسیار عظیم می‌باشند و مدارک بسیار زیادی را پوشش می‌دهند، موتورهای جستجو در اغلب جستجوها هزاران و میلیونها صفحه وب را بازیابی می‌کنند. به همین دلیل برای استفاده مؤثر از موتورهای جستجو بکارگیری فتنونی که با محدود نمودن جستجو تعداد نتایج حاصل از جستجو را کاهش دهد و به این ترتیب موجب بازیابی مدارک مرتبط‌تر با موضوع مورد جستجو شوند، ضروری است.

یکسان نبودن مجموعه‌های موتورهای جستجوی مختلف از یک طرف و تفاوت قواعد و شیوه‌های انجام جستجو و به عبارت دیگر تفاوت الگوریتم جستجو در موتورهای مختلف از طرف دیگر

باعث می‌شوند نتایج بازیابی شده برای جستجوی یک موضوع در همه موتورهای جستجو یکسان نباشند. همچنین این تفاوت‌ها باعث می‌شوند درجه ارتباط هر صفحه وب با یک موضوع مورد جستجوی واحد در چند موتور جستجو با هم یکسان نباشند. بنابراین ممکن است به هنگام جستجوی کلیدواژه‌های دربرگیرنده موضوع واحدی در موتورهای جستجوی مختلف، یک صفحه وب در یکی از موتورها کاملاً مرتبط تعیین شود و همان صفحه وب در موتور جستجوی دیگری دارای ارتباط موضوعی کمتری تشخیص داده شود.

بطور کلی موتورهای جستجو برای بازیابی مدارک معمولاً چند عامل اصلی زیر را مورد توجه قرار می‌دهند: بسامد کلیدواژه، محل قرار گرفتن کلیدواژه، پیوندهای منتهی به مدرک، خوشه‌های^{۲۱} کلیدواژه‌ها (منظور در کنارهم و یا نزدیک به هم قرار گرفتن کلیدواژه‌های مورد جستجو در متن مدرک است).

در مجموع می‌توان گفت تقریباً تمام موتورهای جستجو برای طبقه‌بندی موضوعی مدارک و تعیین درجه ارتباط موضوعی آنها با مبحث مورد جستجو، از شیوه محل / بسامد^{۲۲} استفاده می‌کنند. یعنی اغلب موتورهای جستجو در وهله نخست برای تعیین درجه ارتباط مدارک با موضوع مورد جستجو بسامد کلیدواژه‌های مورد جستجو و محل قرار گرفتن آنها در مدرک را مورد توجه قرار می‌دهند. معمولاً موتورهای جستجو برای شناسایی و بازیابی مدارک، یکی از این دو شیوه متداول را به کار می‌گیرند: جستجوی کلیدواژه‌ای و جستجوی مفهومی^{۲۳}. جستجوی کلیدواژه‌ای از بین این دو شیوه، روش متداول‌تر جستجو در موتورهای جستجو می‌باشد و اغلب موتورهای جستجو برای بازیابی مدارک اینترنتی از روش جستجوی کلیدواژه‌ای استفاده می‌کنند (۱).



در جستجوی کلیدواژه‌های مدارکی بازیابی می‌شوند که حاوی عین واژه‌های مورد جستجو باشند. معیار دفعات تکرار واژه در متن همیشه نمی‌تواند ملاک مناسبی برای تعیین درجه ارتباط متن با موضوع مورد نظر باشد. برای مثال در صورتیکه کلیدواژه شما بیانگر یک مفهوم باشد، برای تشخیص مرتبط بودن اطلاعات محتوای مدرک با موضوع مورد جستجو، نیازی به تکرار این نوع کلیدواژه‌ها در متن نخواهد بود. از طرف دیگر چنانچه کلیدواژه مورد جستجو، واژه‌ای عام، پرستفاده و متداول در متون باشد، و یا واژه‌ای باشد که دارای مفاهیم متفاوت است، مدارک نامرتبط بسیار زیادی در نتیجه چنین جستجویی بازیابی خواهند شد. در جستجوی کلیدواژه‌های مشکلاتی از قبیل موارد زیر می‌تواند بروز کند:

- موتورهای جستجو به هنگام بازیابی کلیدواژه نمی‌توانند کلماتی را که دارای املاهای مشابه ولی معانی متفاوتی می‌باشند از هم تفکیک نمایند.

- موتورهای جستجو نمی‌توانند کلمات مترادف دارای املاهای متفاوت را تشخیص داده و بازیابی کنند، مثل Heart و Cardiac.

در جستجوی مفهومی مدارکی بازیابی می‌شوند که مرتبط با مفهوم مورد جستجو تشخیص داده شوند. در این شیوه ممکن است مدارک بازیابی شده حاوی عین واژه یا واژه‌های مورد جستجو نباشند (۲).

جستجوی مفهومی در موتورهای جستجو همچنین به عنوان خوشه‌ای کردن و یا دسته‌بندی نمودن کلیدواژه‌ها نیز شناخته شده است. برخی از موتورهای جستجو مفهوم نهایی هر کلیدواژه را با توجه به واژه‌های اطراف و نزدیک به آن تعیین می‌کنند. شیوه‌های مختلفی برای این نوع دسته‌بندی مفهومی واژه‌ها وجود دارد، بعضی از این شیوه‌ها بسیار پیچیده و بر اساس نظریه هوش و زبان مصنوعی طراحی

شده‌اند. برای مثال موتور جستجوی Excite برای دسته‌بندی مفهومی واژه‌ها از یک شیوه ریاضی استفاده می‌کند. نرم‌افزار این موتور جستجو مفهوم هر بخش از متن مدرک را از طریق محاسبه تناوب و دفعات تکرار کلیدواژه‌ها در کنار هم و یا نزدیک به هم تعیین می‌کند. اگر چندین واژه که برای بیان یک مفهوم با برچسب خوردن مشخص شده‌اند در متنی نزدیک به هم قرار گرفته باشند، موتور جستجوی Excite از طریق تجزیه و تحلیل آماری نتیجه‌گیری می‌کند که آن بخش متن درباره موضوع خاصی است. بطور مثال زمانی که کلمه قلب در متون پزشکی و بهداشتی بکار می‌رود، معمولاً در کنار کلماتی از قبیل سرخرگ، شش، خون، سیاهرگ، آنورت و غیره قرار می‌گیرد. در صورتی که اگر همین کلمه قلب در کنار واژه‌هایی مثل گل، دوست داشتن و غیره قرار گیرد، موتور جستجو مفهوم متن را درباره عشق تشخیص خواهد داد (۳).

نکته مهم اینکه شیوه و روش فوق بیشتر جنبه نظری دارد تا عملی و کاربردی. جستجوی مفهومی با وجود اینکه شیوه بسیار مفیدی است، در عمل بطور کامل و دقیق قابل اجرا نمی‌باشد. این شیوه زمانی از کارایی بیشتری برخوردار خواهد بود که به هنگام تنظیم عبارت جستجو کلمات زیادی که همه آنها در مجموع به یک مفهوم اشاره دارند، در عبارت جستجو قرار داده شوند. حتی در بهترین شرایط جستجوی مفهومی نیز تنها امکان بازیابی مدارکی که نزدیکی تقریبی و نه تطابق دقیق با مفهوم مورد نظر شما را داشته باشند وجود دارد. در حال حاضر Excite شناخته‌شده‌ترین موتور جستجوی عمومی است که تأکید بر جستجوی مفهومی دارد.

دقت^{۲۴}، فراخوان^{۲۵} و پوشش^{۲۶}، سه مورد از عوامل محدود کننده اغلب موتورهای جستجو هستند (۴). دقت میزان ارتباط موضوعی مدارک بازیابی شده را



Alta Vista	http://www.altavista.com
Answers	http://www.answers.com.au
Euroseek	http://www.euroseek.com
Excite	http://www.excite.com
Fast Search	http://www.hotbot.com
Google	http://www.google.com
HotBot	http://www.hotbot.com
Lycos	http://www.lycos.com
Northern Light	http://www.northernlight.com
Web Wombat	http://www.webwombat.com

ابرموتورهای جستجو:

ابرموتورهای جستجو سرویس‌های جستجوی غول‌آسایی هستند که قادرند بطور همزمان و بسیار سریع اطلاعات مورد نظر را در چند موتور جستجو و فهرست موضوعی بازیابی کنند.

ابرموتورهای جستجو مانند موتورهای جستجو از طریق خزیدن در صفحات وب، و یا تشکیل و نگهداری پایگاه اطلاعاتی متشکل از صفحات وب، موضوعات و کلیدواژه‌های مورد نظر را جستجو نمی‌کنند، بلکه با انتقال موارد جستجو به موتورهای جستجوی اصلی و دریافت نتایج بازیابی شده از این موتورها، نقش واسطه را ایفا می‌کنند (۱).

از آنجا که موتورهای جستجو اغلب تعداد بسیار زیادی مدارک اطلاعاتی متنوع را به عنوان نتایج حاصل از یک جستجو ارائه می‌کنند، ابرموتورهای جستجو قادرند با استفاده از مکانیسمی خاص و روشی سریع، موتورهای جستجویی که توانایی ارائه بهترین نتایج را برای موضوع مورد جستجوی شما دارند را تعیین و معرفی نمایند.

برخی از ویژگی‌های ابرموتورهای جستجو:

- ابرموتورهای جستجو سریع و تقریباً سطحی

با موضوع مورد جستجو می‌سنجد. فراخوان یا جامعیت درصد مدارک مرتبط بازیابی شده را نسبت به مجموع مدارک بازیابی شده می‌سنجد. پوشش نیز اشاره دارد به اینکه چه درصدی از کل مدارک مرتبط بالقوه موجود در اینترنت جزء مجموعه موتور جستجو می‌باشند.

از آنجا که بازیابی مدارک نامرتبط و کم ارتباط با موضوع مورد نظر از مشکلات شایع جستجو در اینترنت است، دقت یکی از مهمترین فاکتورها در هر موتور جستجو محسوب می‌شود. پایین بودن میزان دقت^۲ بازیابی مدارک توسط موتورهای جستجو به دلایل زیر می‌باشد:

۱- بی‌دقتی جستجوگر در تنظیم عبارت جستجو.

۲- اشتباهات موتور جستجو در نمایه‌سازی مدارک اینترنتی.

۳- علامتگذاری کلیدواژه‌هایی که بیانگر محتوای واقعی مدرک نمی‌باشند توسط تهیه‌کننده صفحه وب.

موتورهای جستجو دارای ویژگی‌هایی می‌باشند که برخی از آنها عبارتند از:

- موتورهای جستجو می‌توانند تمام متن صفحات وب برگزیده و منتخب خود را ایندکس کنند.

- موتورهای جستجو اغلب مدارک مورد نظر را از طریق جستجوی کلیدواژه‌ای یا به عبارت دیگر جستجوی عین واژه‌ها و یا عبارت‌های مورد نظر بازیابی می‌کنند.

- در موتورهای جستجو معمولاً امکان مرور موضوعات و جستجوی مرحله‌ای موضوع (جستجوی درختی) نظیر فهرستهای موضوعی وجود ندارد.

اسامی و آدرس‌های اینترنتی برخی از موتورهای جستجوی مهم عبارتند از:



درجه‌بندی این ابزارها، مورد توجه قرار می‌گیرد. بزرگترین ابزارهای جستجو در تاریخ ششم آوریل ۲۰۰۱ (۱۷ فروردین ۱۳۸۰) به ترتیب عبارت بوده‌اند (۱) از:

۱. Google با نمایه‌سازی ۷۰۵ میلیون صفحه وب.
۲. FAST با نمایه‌سازی ۵۷۵ میلیون صفحه وب.
۳. Alta Vista با نمایه‌سازی ۵۵۰ میلیون صفحه وب.
۴. Inktomi با نمایه‌سازی ۵۰۰ میلیون صفحه وب.
۵. WebTop.com با نمایه‌سازی ۵۰۰ میلیون صفحه وب.

پرمراجعه‌ترین ابزارهای جستجو:

دفعات مراجعه به هر سایت اینترنتی، یکی دیگر از عواملی است که به هنگام بررسی کیفیت ارائه خدمات و میزان موفقیت این سایت‌ها مورد توجه قرار می‌گیرد. پرمراجعه‌ترین سایت‌های اینترنتی در تاریخ ۲۶ آگوست ۲۰۰۱ (۴ شهریور ۱۳۸۰) به ترتیب عبارت بوده‌اند (۱) از:

۱. AOL Time Warner با ۳۷۸۴۹۴۱۱ مراجعه.
۲. Yahoo با ۳۰۳۷۹۴۶۵ مراجعه.
۳. MSN با ۲۹۰۹۶۲۶۵ مراجعه.
۴. Microsoft با ۹۸۶۶۴۰۸ مراجعه.
۵. Lycos Network با ۸۲۲۹۹۷۰ مراجعه.

جذاب‌ترین سایتها:

از جمله عوامل دیگر قابل تأمل در ارزیابی سایت‌های اینترنتی، مدت زمانی است که کاربران بطور متوسط صرف استفاده از سایت می‌کنند. به این ترتیب بر اساس بررسی انجام یافته در تاریخ ۲۶ آگوست ۲۰۰۱ (۴ شهریور ۱۳۸۰)، جذاب‌ترین سایت‌های اینترنتی به ترتیب عبارت بوده‌اند (۱) از:

تعدادی موتور جستجو را بطور همزمان مورد جستجو قرار داده و مرور می‌کنند. نتایج حاصل از جستجو را نیز بصورت فهرست‌هایی ارائه می‌نمایند.

-ابرموتورهای جستجو تقریباً تنها ۱۰ درصد نتایج حاصل از جستجوی هر یک از موتورهای جستجو را می‌توانند بازیابی کنند.

اسامی و آدرس‌های اینترنتی برخی از ابرموتورهای جستجوی مهم عبارتند از:

- Chubba: <http://www.chubba.whatuseek.com>
- Copernic: <http://www.copernic.com>
- Dogpile: <http://www.dogpile.com>
- Highway61: <http://www.highway61.com>
- InferenceFind: <http://www.infind.com>
- InfoSearch 2000: <http://www.infosearch2000.com>
- Ixquick: <http://www.ixquick.com>
- Meta Crawler: <http://www.metacrawler.com>
- ProFusion: <http://www.profusion.com>
- Query Server: <http://www.queryserver.com>
- Search Caddy: <http://www.searchcaddy.com>
- Seek 123: <http://www.seek123.com>
- Vivisimo: <http://www.searchenginecolossus.com>
- Web Ferret: <http://www.ferretsoft.com/netferret>

برترین‌های ابزارهای جستجو

در این بخش نتایج حاصل از بررسی‌های انجام یافته در جهت شناسایی برترین ابزارهای جستجو از جنبه‌ها و دیدگاه‌های مختلف ارائه می‌شود.

بزرگترین ابزارهای جستجو

اندازه مجموعه ابزارهای جستجو یکی از اولین و متداولترین ویژگی‌هایی است که معمولاً در



<<http://www.home.sprintmail.com/~debflanagan/Subject.htm>>

4- "A Helpful Guide to Web Search Engines". Monash Information Services. 6 August 1999. [online]. <<http://www.monash.com/spidap4.html>>

5- "How Much Information?". School of Information Management and Systems (SIMS). 2000. [online]. <<http://www.sims.berkeley.edu/how-much-info/charts/charts.html>>. [24 Dec. 2000].

6- "How Search Engines Rank Web Pages". Search Engine Watch. 1999. [online]. <<http://www.searchenginewatch.com/webmasters/rank.html>>

7- "How Search Engines Work". Search Engine Watch. 1998. [online]. <http://www.searchenginewatch.com/webmasters/work.html>>

8- "Important Things to know Before You Begin Searching the Web". University of California Berkeley Library. 14 August 1999. [online]. <<http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/ThingToKnow.html>>

9- "Internet Search Basics and Why There's a Problem". 1 July 1999. [online]. <<http://www.thewebtools.com/tutorial/part2.htm>>

10- "Search Engines, Crawlers, and Spiders". Northwest Educational Technology Consortium. 10 May 1999. [online]. <<http://www.netc.org/basics/engines.html>>

11- "Searches Per Day". Search Engine Watch. 2001. [online]. <<http://www.searchenginewatch.com/reports/perday.html>>. [3 July 2001].

12- "Searching With Internet Provided Resources". 1 July 1999. [online]. <<http://www.thewebtools.com/tutorial/tutorial.htm>>

13- Sullivan, Danny. "Search Engine Sizes". Search Engine Watch 2001. [online]. <http://www.searchenginewatch.com/reports/sizes.html>>. [3 July 2001].

14- "Top 25 Web Properties". Nielsen//NetRatings. 2001. [online]. <<http://pm.netratings.com/nnpm/owa/Nrpublicreports.toppropertiesweekly>>. [1 sep. 2001].

15- "What is a Search Engine". Tripod. 1999. [online]. <http://www.members.tripod.com/WilliamTan/s_engine.htm>

یادداشت‌ها

1- Web Pages

1. Online A با متوسط زمان ۵۸ دقیقه و ۵۲ ثانیه برای استفاده هر کاربر.
2. eBay با متوسط زمان ۳۹ دقیقه و ۴۴ ثانیه برای استفاده هر کاربر.
3. Yahoo با متوسط زمان ۲۸ دقیقه و ۳۲ ثانیه برای استفاده هر کاربر.
4. iWon با متوسط زمان ۲۶ دقیقه و ۱۵ ثانیه برای استفاده هر کاربر.
5. MSN با متوسط زمان ۲۲ دقیقه و ۲۸ ثانیه برای استفاده هر کاربر.

فعالترین ابزارهای جستجو

تعداد جستجوهای انجام گرفته توسط موتورهای جستجو، فهرست‌های موضوعی و ... در هر روز، از دیگر اطلاعات آماری است که در ارزیابی این ابزارهای جستجو، می‌تواند از اهمیت ویژه‌ای برخوردار باشد. به این ترتیب بر اساس گزارش سایت Search Engine Watch فعالترین ابزارهای جستجو به ترتیب عبارتند (۱) از:

1. Google با ۱۰۰ میلیون جستجو در روز.
2. Alta Vista با ۵۰ میلیون جستجو در روز.
3. Inktomi با ۴۷ میلیون جستجو در روز.
4. Direct Hit با ۲۰ میلیون جستجو در روز.
5. FAST با ۱۲ میلیون جستجو در روز.

منابع

- 1- Flanagan, Debbie. "Meta-search Engines". 1999. [online]. <<http://www.home.sprintmail.com/~debflanagan/parallel.htm>>
- 2- Flanagan, Debbie. "Search Engines". 1999. [online]. <<http://www.home.sprintmail.com/~debflanagan/engines.htm>>
- 3- Flanagan, Debbie. "Subject Directories". 1999. [online].