فهرست مطالب :

موتورهای جستجو

مقدمه

رشد اينترنت شگفت‌آور شده است. با توجه به تحقيق ميداني در سال ۱۹۹٦ اين رشد تصاعدي همچنان ادامه دارد. و تخمين زده شده كه شبكه از نظر اندازه و حجم هر ۱۲ تا ۱۵ ماه دوبرابر مي‌شود. بطور تقريبي ۱۰۰/۰۰۰ وب‌گاه در اگوست ۱۹۹۵ وجود داشته و اين تعداد در اگوست ۱۹۹٦ به ۵۳٦/۰٤۱ رسيده است. از آنجائي كه هر پايگاه مي‌تواند بسياري از صفحات وب را در خود داشته باشد اين باعث مي‌شود كه تعداد بيشتري از صفحات وب بوجود آيد. در حاليكه كيفيت بسياري از صفحات ممكن است مورد سؤال باشد و نگهداري بسياري از صفحات پراكنده است. اما صفحات معتبري هم وجود دارد كه اطلاعات با ارزشي در مورد بسياري از موضوعات ارائه مي‌دهد. آنچه استفاده‌كنندگان به آن نياز دارند يك سيستم جداكننده است كه مو از ماست بكشد.

تلاشهاي زيادي صورت گرفته كه اين وظيفه را آسان كند: كتابداران و متخصصان موضوعي راهنماهاي موضوعي را گردآوري كرده‌اند.«اخبار كتابخانه‌هاي تحقيقاتي و دانشكده‌اي»۱ بطور دوره‌اي راهنماهاي منابع اينترنتي را در موضوعات انتخابي منتشر مي‌كند. علاوه بر آن فهرستنويسان Intercat را (كه يك فهرست آزمايشي براي اينترنت است) بوجود آورده‌اند. را براي نمايه‌سازي منابع اينترنت با سرعنوانهاي موضوعي Netfirst نيز پايگاه OCLC كتابخانه كنگره و اعداد طرح دهدهي ديويي ايجاد كرده است. كتابداران مرجع يك پايگاه وب را بوجود آورده‌اند كه مرور منابع اينترنتي را به اشتراك بگذارند و كتابداران Infofilterبه نام رده‌بندي موضوعي را براي سازماندهي منابع اينترنت مورد آزمايش قرار داده‌اند. اما چيزي كه بيشترين توجه استفاده‌كنندگان اينترنت را به خود معطوف داشته است، احتمالاً عنكبوتها و روباتهايي هستند كه خدمات جستجو را انتخاب مي‌كنند. براي بسياري از جستجوگران اينترنت، اين موتورها با راه دادن آنها به فضاي اطلاعاتي عظيم كمك موقتي ارائه مي‌دهند. كاربران اينترنت بزودي فهميدند كه اين موتورها كامل و مناسب نيستند زيرا آنها منطقه جغرافيايي خاصي را پوشش مي‌دهند، بصورت متفاوت نمايه‌سازي مي‌شوند و منابع را با كليدواژه‌ها بازيابي مي‌كنند. استفاده‌كنندگان هرگز نمي‌توانند اعتماد كنند كه يك جستجو جامع يا قطعي باشد. با وجود اينكه نتايج بازيابي ظاهراً بوسيله ميزان ارتباط مرتب شده است و استفاده‌كنندگان مبارزه

با ريزش كاذب مواد تكراري و نامربوط را ادامه مي‌دهند. در كل پيشرفت خوبي در كمك به استفاده‌كنندگان براي جهت‌يابي در اينترنت بوجود آمده، اما اين ابزارها آنقدر زياد هستند كه انتخاب صحيح يكي از آنها كار سختي است.

علاوه بر آن انتخاب موتورهاي جستجوي مناسب ممكن است براي استفاده‌كنندگان عمومي و نيز كتابداران، به علت رويه‌هاي پيچيده، مهمتر باشد. بعنوان مثال، اينفوسيك سرويس رايگان دارد، اما سرويسهاي هزينه بر، يعني متخصصان اينفوسيك، نمايهء بزرگتر و قابليتهاي جستجوي قويتر ارائه مي‌دهند.

وب كروكر يك درخواست حق عضويت دارد كه زماني آن را تحميل مي‌كند. اگر اين عمل در مقابل هزينه براي خدمات يك رويه شود، لازم خواهد بود براي استفاده كنندگان و بخصوص كتابداران كه بدانند كداميك از موتورهاي جستجو را بايد به خدمت بگيرند.

اين تحقيق تلاش كرده كه كارآيي موتورهاي كاوش را در آدرس دهي نيازهاي اطلاعاتي ارزيابي كند. آيا آنها مي‌توانند جوابهايي براي سؤالات مرجع واقعي بازيابي كنند؟ آيا آنها منابع خوبي براي سؤالات موضوعي ارائه مي‌دهند؟ آنها تا چه حد نتايج جستجو را براساس ميزان ارتباط مرتب مي‌كنند؟ كداميك از موتورهاي جستجو بهتر عمل مي‌كنند؟ جواب اين سؤالات به ما كمك خواهد كرد كه نقاط ضعف و قوت موتورهاي كاوش را بهتر بفهميم و ما را قادر مي‌سازد كه براي برطرف كردن نيازهاي اطلاعاتي موتور كاوش مناسب را انتخاب كنيم.

پيشينه پژوهش

Netsearchموتورهاي كاوش بسياري موجود هستند و براحتي در دسترس قرار مي‌گيرند مربوط به كتابخانه كنگره، هر دو با هم www by subject or keywordمتعلق به نت اسكيپ و موتورهاي كاوش اصلي را گرد آورده‌اند. همچنين ابرموتورهايي وجود دارد كه به جستجوگران قادر Savy searchاجازه مي‌دهد كه سريعاً به چندين موتور كاوش دسترسي پيدا كنند، اما فقط گروهي ديگر از ابزارهائي را "All-in-oneبه جستجوي بيش از ٥ پايگاه در يك زمان است. " كه ارائه دهنده نمونه‌هاي جستجو بسياري از موتورهاي كاوش در يك پايگاه وب براي آسان‌سازي عمل جستجو هستند را نشان مي‌دهد.

كار عنكبوتها، روباتها و ديگر برنامه‌هاي خودكار بوسيله پروسيس خلاصه شده است. (١٩٩٥)

و محدوديتهاي اين ابزارها بوسيلهء كاستر تجزيه و تحليل شده‌اند. (١٩٩٥) چندين مقاله ادعا كرده‌اند كه ابزارهاي جستجوي اينترنت را ارزيابي كرده‌اند، اما اغلب آنها توصيفهايي از شبكه جهاني وب يا موتورهاي كاوش ارائه داده‌اند. برينكلي و بيورك (١٩٩٥) هايتلنت، آرچي، گوفر، و ايزو شبكه جهاني وب را شرح داده‌اند. كورتوليس، بيرواستارك (١٩٩٥) پرسشهايي بكار برده‌اند براي اينكه موتورهاي كاوش و نمايه‌هاي وب را آزمايش كنند. اما گزارش آنها بيشتر توصيفي است. كايمل (١٩٩٦) تاريخچه‌اي از پايگاههاي توليد شده بوسيله رباتها را ارائه داده و به جستجوگران مبتدي راهنمائي‌هاي خوبي در مورد موتورهاي كاوش پيشنهاد كرده است. وندیتو(١٩٩٦) هفت موتور كاوش را آزمايش كرده و ويژگيهاي جستجوي آنها را شرح داده است. گزارشهاي مشابه در مورد اينترنت فراوان است و بسياري از آنها را مي‌توان در ليت كمپبل پيدا كرد. نمونه‌هاي خوب بسيار كم هستند. مثل گزارشهاي لين (١٩٩٥)، وين شيپ (١٩٩٥) وبستر و پانول (١٩٩٥).

مطالعات ارزشيابي نسبتاً كمي وجود داشته كه شامل تحقيق ميداني در مورد موتورهاي جستجو باشد. دساي (١٩٩٥) قدرت بازيابي سيزده ابزار جستجو را با يك سؤال آزمايش كرده است. او قادر بود با جستجوي نامش تعيين كند كه چگونه بسياري از اسناد وبي او بازيابي شده است. اينفوسيك و لايكاس با بازيابي هفت سند از ٢٤ سند بهتر عمل كردند. در حاليكه ديگر موتورها و نمايه‌ها نظير وب كرولر و ياهو ضعيف عمل كردند. ليتون (١٩٩٥) عملكرد اينفوسيك، لايكاس، وب كرولر و ورلد وايد وب وارم را با استفاده از ٨ سؤال مقايسه كرده است. او اين آزمايش را با ٤ معيار انجام داد-نسبت تكراري بودن، دقت، دقت كامل و حداكثر ١٠ دقت- و نتيجه گرفت كه لايكاس و اينفوسيك بهتر از بقيه عمل كردند. پكروتومايولو ٢ سؤال مرجع را در آلتاويستا، ماژلان، اينفوسيك، لايكاس و يونيت جستجو كردند. امتياز دقت آنها مبتني بود بر ١٠ نتيجه اول، آنها فهميدند كه آلتاويستا بهترين عملكرد را داشت بعد از آن اينفوسيك، لايكاس، ماژلان و پوينت. مقابقاب (١٩٩٥) ٥ سؤال را براي امتحان كردن ٥ موتور كاوش با اجراي جستجوهاي اصلي و اصلاح شده در هر موتور بكار برد. او دقت نتيجه اول را تغيير داد. با استفاده از ٢٥ نتيجه بعنوان پايه و مبنا. او ياهو را بعنوان بهترين عمل كننده شناخت. يافته‌هاي اين مطالعات بطور قطعي بهترين موتورهاي كاوش را مشخص نكرد بعلت سؤالات مختلف، تعداد متفاوت سؤالات و مقياسهاي مختلف بكار رفته براي ارزشيابي. با اين وجود، اين مطالعات شيوه‌هاي مختلف

ارزيابي موتورهاي كاوش را ثابت كرد و معيارهاي جديد و منطقي براي اجراي جستجو پيشنهاد كرد.

روش‌شناسي

اين مطالعه هشت موتور كاوش را كه مشهور هستند و براي عموم رايگانند ارزيابي مي‌كند. اين موتورها عبارتند از: آلتاويستا، اكسايت، اينفوسيك، گايد، لايكاس، ماژلان، اين تكست، وب كرولر و ورد وايد وب وارم.

عوامل بسياري ممكن است براي موفقيت يك جستجو مشاركت كنند. درك صحيح از تقاضاي جستجو، استراتژي جستجو، پايگاه اطلاعاتي، موتور كاوش، و قضاوت ميزان ارتباط بوسيلهء جستجوگران. در اين تحقيق تعدادي از اين معيارها كنترل شده بودند بنحوي كه تفاوت موتورهاي كاوش قابل مشاهده بود. موتورهاي كاوش از نظر اندازه، محتواي پايگاه اطلاعاتي‌شان، خطمشي‌هاي نمايه‌سازيشان، كنترل كيفيت، شيوه‌هاي بازيابي و ارائه نتايج جستجو متفاوتند. فرض بر اين شد كه سؤالات آزمايشي با پيچيدگي و وضع متفاوت در تعيين بهترين موتور جستجو با ارزش‌تر خواهد بود. با اين وجود، مطالعات پيشين مشخص كردند كه امكان اينكه يك موتور جستجو در جوابگويي همه نوع سؤالات بهتر از همه باشد، وجود ندارد. ما تقاضاهاي جستجو را با استفاده از ۲۰ سؤال مطرح شده در ميز مرجع استاندارد كرديم. و ۵ سؤال موضوعي كه در حوزه‌هايي كه منابع اينترنتي بسياري داشت، بوجود آمده بود-سرگرمي، تجارت، سياست اقتصاد و بهداشت. سؤالات مرجع گردآوري شده شامل سؤالات تخصصي و پرسشهاي موضوعي وسيع بود و از نظر اينكه توانايي موتورهاي كاوش را در جوابگويي به سؤالات مرجع واقعي مورد آزمايش قرار داد با ارزش بودند. ۵ سؤال موضوعي ساختگي بودند، اما اين طراحي ما را قادر مي‌ساخت كه تجزيه و تحليل معني‌دار بيشتري انجام دهيم. با اين همه، هيچ ارزشيابي خيلي خردمندانه نبود اگر ما از سؤالاتي كه براي آن هيچ چيز قابل بازيابي نبود، استفاده مي‌كرديم.

مجموعه داده‌ها۲ به هشت موتور كاوش چهار جستجو اختصاص يافته بود براي اطمينان از اينكه

هر سؤال دو بار در يك موتور جستجو شده است. به جستجوگران آموزش داده شد كه از مرورگر نت اسكيپ براي دسترسي به اينترنت استفاده كنند و سؤالات داده شده را در موتورهاي كاوش تعيين شده جستجو كنند و نسخه‌هاي چاپي از نتايج جستجو ارائه دهند. جستجوگران كار را در آوريل شروع كردند و در ژوئن ١٩٩٦ نتيجه‌گيري انجام شد. جمله‌بندي سؤالات مرجع براي جستجوهاي اينترنتي كمي تغيير داده شد. براي اطمينان يافتن از تكنيكهاي مشابه بكار رفته، كليدواژه‌ها تعيين شده بودند و پارامترهاي اساسي در مورد اينكه در هر موتور كاوش جستجو به چه صورت انجام شود ارائه شدند. به جستجوگران گفته شد كه بهترين قضاوتشان را در ارزيابي ارتباط منابع بازيابي شده بكار گيرند.

يكي از سؤالات مرجع دو بخش داشت، بنابراين سئوالات مرجع به ٢١ تبديل شد. هر كداميك از سؤالات ٢١ گانه مرجع و ٥ سؤال موضوعي دو بار در هر موتور، جستجو شده بود. اما در اكسايت ٤ بار جستجو شد، زيرا در اين موتور، جستجوگر قادر به جستجوي كليدواژه‌اي و نيز جستجوي مفهومي مي‌باشد. در ميانه راه با ماژلان جستجوگران گزينه‌اي از جستجو را در بخش خاصي از پايگاههاي اطلاعاتي يا كل پايگاهها به منظور هماهنگي و يكدستي همه جستجوگران در كل پايگاهها انجام دادند. در كل ٤٦٨ جستجو انجام شده بود.

متغيرهاي وابسته. چهار متغير براي اين تحقيق اندازه‌گيري شدند. «دقت»٣ كه بطور سنتي تعريف شده بود: تعداد منابع مرتبط بازيابي شده تقسيم بر تعداد منابع بازيابي شده و يك معيار استانداردي براي سيستمهاي بازيابي اطلاعات بوده است. از آنجائي كه ارزيابي ارتباط تعداد زياد صفحات بازيابي شده بوسيلهء موتورهاي كاوش غيرممكن بود .

اين متغير در اين تحقيق بطور عملياتي چنين تعريف شده:

دقت: تعداد منابع مرتبط در ١٠ گزينه اول

شيوه استفاده از ده گزينه اول قابل توصيه است، زيرا اين گزينه‌ها بيشتر امكان دارد كه به وسيله جستجوگران ديده شود. اين معيار را ليتون، پيكروتومايولو بكار برده است. اما برخلاف تحقيق ليتون، اين تحقيق پايگاههاي ارجاعي و تكراري را در معيار دقت درنظر مي‌گيرد. زيرا آنها بالقوه مفيد بودند (در صورتي كه گزينه‌هاي تكراري مرتبط باشد) و حذف آنها باعث مي‌شد كه پايهء مقايسه (كه ده تا بود) كوچكتر شود.

موتورهای جستجو

«تكراري بودن»٤: در همان اوايل جستجو در موتورهاي كاوش گزارشهاي حكايت گونه‌اي از تكراريها بوجود آمد. در نتيجه اين معيار در ارزيابي‌هاي ما وارد شد. تكراري بودن بطور عملياتي «تعداد گزينه‌هايي كه تكرار شدند و قبل از آن نيز ارائه شده بودند» تعريف شده بود. پايگاههاي ارجاعي هم جزء تكراريها به حساب آمدند. در اين مطالعه ما تعداد گزينه‌هاي تكراري را بر اساس ده نتيجه اول در نظر گرفتيم.

«امتياز مرتبطترين گزينه»٥: همه موتورهاي كاوش انتخابي، نتايج بازيابي را با استفاده از الگوريتم متفاوت مرتب مي‌كنند و بهترين تطبيقها را اول ارائه مي‌دهند. اما تنظيم هميشه مفيد نبوده است. اين متغير براي امتحان كردن توانايي درجه‌بندي موتورهاي كاوش طراحي شده بود. كه بر اين فرضيه مبتني است كه شيوه درجه‌بندي مؤثر، مرتبطترين گزينه‌ها را در بالاترين ليست نتايج جستجو قرار مي‌دهد. جستجوگران، بطور عملياتي مرتبطترين گزينه از بين ده گزينه تعريف كردند و به آن بخاطر جايگاهش يك امتياز دادند. اگر اين گزينه در اولين، دومين يا سومين گزينه بود اين موتور امتيازي بين ١ يا ٢ يا سه مي‌گرفت. اگر اين گزينه جاي ديگر ظاهر مي‌شد به آن امتياز ٦ داده مي‌شد. عدد ٦ به اين علت انتخاب شده بود كه نشان مي‌داد اين گزينه در خارج از اولين نيمه ليست ده تايي قرار گرفته است. امتياز پائين در مورد مرتبطترين گزينه نشان دهنده اين بود كه آن موتور بهترين درجه‌بندي گزينه‌هاي مرتبط را داشته است.

«امتياز درجه‌بندي ميزان ارتباط»٦: اين متغير نيز درجه‌بندي ميزان ارتباط در موتورهاي كاوش را ارزيابي كرد اما به شيوه‌اي متفاوت. اين متغير به عنوان درصد گزينه‌هاي مرتبط كه در اولين نيمه ليست ده گزينه‌اي ظاهر شدند تعريف شد. اين تعريف مبتني بود بر اين فرضيه كه ميزان ارتباط گزينه‌ها كاهش خواهد يافت هر چقدر كه به گزينه‌هاي پائين‌تر مي‌رسيم.

جستجوگران تعداد گزينه‌هاي مرتبط را در هر نيمه از ليست ده گزينه‌اي ثبت كردند و مأموران تحقيق اين تعداد را با توجه به فرمول زير براي رسيدن به امتياز درجه‌بندي براساس ميزان ارتباط تبديل كردند:

تعداد گزينه‌هاي مرتبط در اولين ليست

كل تعداد گزينه‌هاي مرتبط در ليست ده‌تايي

«بازيابي»٧: يكي ديگر از معيارهاي استاندارد براي بازيابي اطلاعات است و چنين تعريف شده: تعداد گزينه‌هاي مرتبط بازيابي شده تقسيم بر كل تعداد گزينه‌هاي مرتبط در يك فايل

اطلاعاتي. اين معيار براي استفاده دشوار بود زيرا جستجوگران مي‌بايست همه گزينه‌هاي مرتبط را در كل يك پايگاه يا فهرست شناسائي كنند. اين مشكل در شبكه جهاني وب خيلي شديدتر است. با هزاران هزار صفحه وب نمايه شده به وسيله موتورهاي جستجو انتخابي غيرممكن بود كه همه صفحات وب مرتبط با موضوع جستجو شناسايي شود. به اين ترتيب بازيابي در اين مطالعه استفاده نشد.

«تجزيه و تحليل داده‌ها»۸: از ٤٦٨ جستجوي انجام شده ٤ معيار براي هر جستجو در موتور كاوش ثبت شد. بسامد و ميانگين اين معيارها براي هر موتور جستجو با نوع سؤالات حساب شده بودند.

يافته‌ها

دقت

سؤالات مرجع عمومي متنوع بود بطوري كه شايد يك كتابدار مرجع براي يافتن پاسخ آنها از اينترنت استفاده نمي‌كرد. با اين وجود همه سؤالات در موتورهاي كاوش جستجو شده بودند كه توانائي‌شان را در پاسخگويي به سؤالات مرجع ارزيابي كنند. موتورهاي كاوش اين كار را بخوبي انجام ندادند. ميانگين امتياز دقت خيلي پائين بود. بين ۰/۳۱ و ۲/۹۳. اين تكست بالاترين تعداد گزينه‌هاي مرتبط را بازيابي كرد. بعد از آن آلتاويستا و اينفوسيك و سپس لايكاس با اختلاف كم چهارم شد. براي نشان دادن جنبه ديگري از اين جستجوها، اطلاعاتي در مورد نقاط كور هر موتور در جدول ۱ قرار گرفتند كه نشان داد اكسايت پائين‌ترين تعداد نقاط كور را داشت و بعد از آن اين تكست و لايكاس. بر رويهم رفته، اين دو مجموعه از داده‌ها اين تكست را بعنوان بهترين موتور در برخورد با سؤالات مرجع معرفي كرد. اين موتور صفحات وب را براي اين سؤالات بازيابي كرد و نتايج جستجويش بالاترين امتياز ميزان دقت را داشت.

در اين پژوهش موتورهاي جستجو با سؤالات موضوعي ساختگي بهتر عمل كردند. سؤالات موضوعي پائين‌ترين ميانگين امتياز دقت (۳/۲) نسبت به بالاترين امتياز ميزان دقت (۲/۹۳) در سؤالات مرجع واقعي بالات بود. اينفوسيك بهتر عمل كرد بعد از آن ماژلان و اين تكست و باز هم لايكاس با اختلاف كم چهارم شد. از آنجائي كه سؤالات مرجع براي حوزه‌هايي طراحي شده

بودند که در مورد آن اطلاعات بیشتري در وب موجود باشد، مشکل نقاط کور در این سؤالات خیلي جدي نبود. در سؤالات مرجع کیفیت گزینه‌هاي بازیابي شده، «دقت خاص» تعداد جستجوهایي که بیش از ٥ گزینه مرتبط را بازیابي کردند در نظر گرفته شد که در جدول ٢ آمده است.

اینفوسیك باز هم برنده ظاهر شد، بعد از آن ماژلان و این تکست و لایکاس و وب کرولر هر سه بطور مساوي در جایگاه سوم قرار گرفتند. این اطلاعات نشان داد که اینفوسیك در برخورد با سؤالات بهترین بود. این موتور بیش از ٥ گزینه مرتبط را براي اغلب پرسشهاي موضوعي بازیابي کرد ونتایج جستجویش بالاترین امتیاز میزان دقت را داشت.

تکراري بودن

تکراریها در بازیابي زمان جستجوگران را تلف مي‌کنند و باعث سردرگمي مي‌شوند. علاوه بر قصه شکایت در مورد تکراریها، به هر حال، این مسأله به نظر مي‌رسد که در بیشتر موتورهاي کاوش مطرح بوده است. میانگین تعداد موارد تکراري براي هر دو سؤالات مرجع و سؤالات موضوعي در هر موتور کاوش ناچیز است (کمتر از یك). اما سؤالات موضوعي شانسشان براي داشتن موارد تکراري بیشتر بود. این اطلاعات نشان مي‌دهد که نمایش داده‌هاي تکراري حتي زماني که گزینه‌هاي مرتبط زیادي بازیابي شده بود اهمیت چنداني نداشتند.

امتیاز مرتبط ترین گزینه

این امتیاز توانایي هر موتور کاوش را براي نشان دادن اولین گزینهء مرتبط اندازه‌گیري کرد. بخاطر اینکه امتیازي به محل گزینه‌ها اختصاص یافته بود، پائین‌ترین امتیازها عملکردهاي بهتر را نشان مي‌داد. براي سؤالات مرجع امتیاز موتورهاي جستجو بین ٣/٣ و ٥/٣ قرار داشت، اول این تکست بعنوان برنده بعد از آن اکسایت و آلتاویستا قرار گرفتند. موتورهاي کاوش با سؤالات موضوعي خوب عمل نکردند. امتیازات آنها بین ٢/٥ تا ٤/٢ قرار داشت. این تکست و بعد از آن اینفوسیك و وب کرولر بهترین عملکرد را داشتند. این تکست در ارائه مرتبطترین گزینه همیشه بهترین بود.

امتیاز تنظیم براساس ارتباط

این امتیاز قدرت موتورهاي کاوش را اندازه‌گیري کرد براي ارائه گزینه‌هاي مرتبط در اولین نیمه نتایج جستجو. براي سؤالات مرجع، امتیازات موتورهاي کاوش بین ١٥/٥% تا ٤٥/١% و

با پيشتازي اين تكست و بعد از آن اينفوسيك و اكسايت قرار داشت. براي سؤالات موضوعي امتيازاتشان بين ٢٣% تا ٥٢/٨% قرار گرفت. اينفوسيك بعنوان بهترين عمل كننده لايكاس در جايگاه دوم و اكسايت به عنوان سومين جايگاه.

عملكرد جامع

چهار معيار جنبه‌هاي قدرت بازيابي موتورهاي كاوش را اندازه‌گيري كردند. نمودار ٢ دقت، تكراري بودن و امتياز مرتبطترين گزينه‌ها را براي سؤالات مرجع خلاصه كرده است. امتياز رتبه‌بندي براساس ميزان ارتباط در آن وارد نشد، زيرا دامنه آنها خيلي بالاتر بود و نمي‌توانست بطور كامل در اين نمودار وارد شود. بهترين موتور جستجو بالاترين دقت، پائين‌ترين موارد تكراري، پائين‌ترين امتياز مرتبط ترين گزينه و بهترين امتياز تنظيم براساس دقت را دارد. اين نتايج در نمودار ٢ روشن است، به هر حال موتورهاي كاوش چنين عمل كردند: اپن تكست بالاترين مانعيت و پائين‌ترين امتياز مرتبطترين گزينه را داشت. اما اكسايت و اينفوسيك پايين‌ترين تعداد موارد تكراري را داشتند. از اين ٤ معيار، اپن تكست بهترين امتياز را از بين آنها داشت و توانست بهترين عمل كننده براي اين نوع سؤالات باشد. رتبه دوم مشخص نبود چون اين موتورها فقط در يك يا دو معيار ممتاز بودند. با اين وجود اين امكان وجود داشت كه آنها را به دو گروه تقسيم كنيم: آلتاويستا، اكسايت، اينفوسيك و لايكاس نسبتاً بهتر از ماژلان، وب كرولر و ورلدواید وب وارم عمل كردند.

نمودار ٣ شباهت دشواري را در تعيين برنده براي سؤالات موضوعي نشان مي‌دهد. اطلاعات موجود بر روي نمودار ٣ اينفوسيك را بهترين عملگر مي‌داند و امتياز تنظيم براساس ارتباط آنرا تقويت مي‌كند. بقيه موارد براي اعلام كردن خيلي مشكل بود.

جستجو در اينترنت

در سال ٢٠٠٠ حدود يكصدميليون پايگاه وب بر روي شبكه جهاني اينترنت وجوددارد و پيش بيني مي شود كه تنها پس از گذشت ٢ سال ، در سال ٢٠٠٢ به ٢٥٠ ميليون پايگاه برسد. با رشد تصاعدي حجم اطلاعات ، يافتن اطلاعات موردنظر در اين درياي پهناور كار مشكلي است و بكارگيري ابزارهاي جستجوي مناسب يكي از ضروريات كار باشبكه مي باشد.

موتورهای جستجو

موتورهاي جستجو از سال ١٩٩٤ مورداستفاده قرار گرفتند. در ابتدا فعاليت آنها فقطجستجو در وب بود ولي با گذشت زمان ، خدمات ديگري از جمله فروش كالا، اجاره فضاي وب و پست الكترونيك ، تحليل سايتها و... به فعاليت آنها اضافه شد.

انواع موتور جستجو

موتورهاي عمومي كه در كليه پايگاهها فارغ از نوع آن جستجو مي كنند.

موتورهاي عمومي معروف عبارتند از:

altavista.com, google.com, go.com, hotbot.com .

"، "deja.com"، تجارت "news.comموتورهاي تخصصي در يك رشته خاص مانند اخبار "

"، مقالات و انتشارات whowhere.com٢- " "،افرادyellow pages.comشركتها "

"infojump.com

موتورهاي تخصصي در خدمات اينترنت deja.com وmagellan .com بهترين پايگاه 3–مانند گروههاي خبري و مباحثه

- ٤- موتورهاي كلان ; اين موتورها عبارت مورد جستجو را همزمان به چند موتورجستجو داده و پاسخها را اولويت بندي كرده و با ذكر نام

- موتور جستجو نمايش مي دهند.مانند: mamma.com,savvysearch.com

٥-نقطه شروع ; اينگونه سايتها موتورهاي جستجو را برحسب موضوعات مختلف معرفي مي كنند. درصورتي كه موتورهاي تخصصي و عمومي رانمي شناسيد از اين پايگاهها شروع كنيد. مانند:

٦-موتور اختصاصي پايگاهها; بعضي از پايگاههاي بزرگ مانند مايكروسافت ، جنرال جديدا مي توان از موتورهاي جستجو الكتريك ،... از داخل پايگاه خود موتور جستجو دارند.

,... در يك پايگاه جهت جستجو در آن استفاده كرد.hotbot,altavistaعمومي مانند

تقسيم بندي ديگري كه از موتور جستجو مي توان كرد، موتورهاي جستجوگر، و دوم موتورهايي

گفته مي web directory كرده اند و به آنها ۷- است كه پايگاهها را دسته بندي موضوعي است . درحال حاضر اكثر موتورهاي جستجو webcrawear,yahooشود. معروفترين آنها دسته بندي موضوعي نيز دارند و در هر دسته و يا زيرشاخه هاي بعدي مي توان جستجو كرد.

نحوه كار موتورهاي جستجو

در اين مقاله فقط نحوه كار موتورهاي عمومي بررسي مي شود. موتورهاي براي يافتن و spider و يا crawler,robotعمومي ازبرنامه هايي معروف به مرور صفحات وب استفاده مي كنند. نحوه كار اين برنامه ها بدين صورت است كه با يافتن يك صفحه كلمات مورداستفاده در آن را شناسايي كرده و به جداول فهرست بانك اطلاعاتي خود اضافه مي كنند درواقع موتورها صفحات وب را در بانك اطلاعاتي نگهداري نمي كنند بلكه دربانك اطلاعاتي فهرستي از كلمات وآدرس صفحات مشمول اين كلمات مي باشد.

كار ديگر اين برنامه ها اين است كه به صفحات فهرست شده قبلي مراجعه كرده ودرصورت به روز شدن صفحات ، مجددا آنها را فهرست بندي مي كنند. ممكن است پايگاه موردنظرموجودنبوده وياآدرس آن عوض شده باشد.

عوامل مهم در فهرست كردن يك صفحه وب ، تعداد وقوع كلمه در صفحه ، محل قرارگيري آن ، نوع فايلهاي مورداستفاده در صفحه ، درجه اهميت كلمه در صفحه با توجه به كليد واژه هاي تعيين شده توسط مالك صفحه و توضيحات آمده در بخش دستورات metaدر شناسنامه صفحه مي باشد.موتورهاي جستجو باتوجه به حجم بانك اطلاعاتي و برنامه هايشان به پايگاههاي جديد مراجعه مي كنند ولي مطمئن تمام پايگاههاي وب را شامل نمي شوند. بزرگترين موتور جستجوي عمومي حدود مي شود. درصورتي كه مي خواهيد پايگاه وب شما ۵۰۰ ميليون صفحه وب را شامل به بانك اطلاعاتي يك موتور جستجو اضافه شود پايگاه وب خود را به آن موتور

جستجو معرفي كنيد در صفحات وب ، در بخش دستورات شناسنامه اي صفحه ، كليد واژه هاي موردنظر خود را معرفي كنيد. موتورهاي جستجوعمومي به دو روش كلمات را فهرست بندي مي كنند.

اغلب موتورهاي جستجو براساس كلمات فهرست بندي مي كنند.درواقع درجستجوكلمات هم keyword indexingمعني راتشخيص نمي دهند

excite.comمعروفترين موتوري كه براساس مفهوم جستجو مي كندمي باشد.

نكته ديگري كه در فهرست بندي بايد بدانيد اين است كه هر موتور چه بخشهايي ازيك صفحه ها ويا چند خط hyperlink,heading, titleرا فهرست مي كند. براي مثال ممكن است فقط مي توان تعيين opentextاول صفحه را فهرست كند. در برخي از موتورهاي جستجو مانند نكته heading,titleكرد كه كلمه مورد جستجو در كجاي صفحه باشد. براي مثال در...,

هستند ما stop wordديگر كلمات معروف به

و... بعضي از موتورها اين كلمات را درنظر نمي web,and,or,the,is,an,a گيرند.مانند

عوامل مهم در انتخاب موتور جستجو

عوامل زيردرانتخاب موتورجستجومهم هستند:

حجم بانك اطلاعاتي موتور جستجو و تعداد صفحات مرور شده توسط آن

به روز بودن بانك اطلاعاتي

تعداد صفحات مرورشده در روز

سرعت برگرداندن نتايج جستجو

تعداد سرويس دهنده هاي آن در شبكه اينترنت جهت كاهش ترافيك و افزايش سرعت

نحوه نمايش نتايج جستجو و كنترل آن توسط كاربر

نحوه اولويت بندي نتايج حاصله و ارتباط آنها با يكديگر و كنترل آن توسط كاربر راحتي استفاده

موتورهای جستجو

صفحات معرفي شده به آن طي چند روز در فهرست قرار مي گيرند

امنيت در پايگاههاimagemap,frameپشتيباني ,

قابليت جستجو در نتايج

stop wordsپشتيباني

حساس به حروف بزرگ و كوچك

پشتيباني عبارت

عدم محدوديت درتعدادحروف عبارت جستجو

دسته بندي موضوعي وامكان جستجودرهردسته

جستجو در خدمات اينترنت شامل وب ، گروههاي خبري و مباحثه ،...

پشتيباني عملگرهاي جستجو

پشتيباني زبانهاي مختلف وتبديل زبانهابه يكديگر

جستجو براساس تاريخ

" bannerارائه امكانات بهتر از جمله ارسال نتايج به آدرس پست الكترونيك ، ارائه تبليغات مرتبط با عبارت .

جستجو براساس نوع فايل مانند تصوير

جستجو در مكان خاصي از صفحه وب

" صفحه وب tagجستجو در دستورات "

ارائه كليد واژه ها و نتايج مشابه

دسته بندي موتور هاي جستجو

موتورهاي جستجو به دو دسته كلي تقسيم مي‌شوند. موتورهاي جستجوي پيمايشي و فهرستهاي تكميل دستي. هر كدام از آنها براي تكميل فهرست خود از روشهاي متفاوتي استفاده ميكنند كه هر يك را بطور جداگانه مورد بررسي قرار مي‌دهيم:

موتور های جستجو

موتورهاي جستجوي پيمايشي يا Crawler-Based Search Engine

ليست خود را بصورت خودكار تشكيل مي‌دهند. Googleموتورهاي جستجوي پيمايشي مانند آنها وب را پيمايش كرده و سپس كاربران آنچه را كه مي‌خواهند از ميانشان جستجو مي‌كنند. اگر شما در صفحه وب خود تغييراتي را اعمال نماييد، موتورهاي جستجوي پيمايشي آنها را به خودي خود مي‌يابند و سپس اين تغييرات ليست خواهند شد. عنوان، متن و ديگر عناصر صفحه، همگي شامل اين ليست خواهند بود.

فهرستهاي تكميل دستي يا Human-Powered Directories

وابسته به كاربراني است كه Dmozمثل Open Directoryيك فهرست تكميل دستي مانند يك آنرا تكميل مي‌كنند. شما صفحه مورد نظر را به همراه توضيحي مختصر در فهرست ثبت مي‌كنيد يا اين كار توسط ويراستارهايي كه براي آن فهرست در نظر گرفته شده انجام مي‌شود. عمل جستجو در اين حالت تنها بر روي توضيحات ثبت شده صورت مي‌گيرد و در صورت تغيير روي صفحه وب، روي فهرست تغييري بوجود نخواهد آورد. چيزهايي كه براي بهبود يك فهرست‌بندي در يك موتور جستجو مفيد هستند، تاثيري بر بهبود فهرست‌بندي يك دايركتوري ندارند. تنها استثناء اين است كه يك سايت خوب با پايگاه داده‌اي با محتواي خوب شانس بيشتري به نسبت يك سايت با پايگاه داده ضعيف دارد.

موتورهاي جستجوي تركيبي با نتايج مختلط

به موتورهايي اطلاق مي‌شود كه هر دو حالت را در كنار هم نمايش مي‌دهند. غالباً، يك موتور جستوي تركيبي در صورت نمايش نتيجه جستجو از هر يك از دسته‌هاي فوق، نتايج حاصل از بيشتر نتايج حاصل از MSNدسته ديگر را هم مورد توجه قرار مي‌دهد. مثلا موتور جستجوي فهرستهاي تكميل دستي را نشان مي‌دهد اما در كنار آن نيم نگاهي هم به نتايج حاصل از جستجوي پيمايشي دارد.

موتور های جستجو

بررسي يك موتور جستجوي پيمايشي

موتورهاي جستجوي پيمايشي شامل سه عنصر اصلي هستند. اولي در اصطلاح عنكبوت است كه پيمايشگر مي‌شود. پيمايشگر همينكه به يك صفحه مي‌رسد، آنرا مي‌خواند و سپس لينكهاي آن به صفحات ديگر را دنبال مي‌نمايد. اين چيزيست كه براي يك سايت پيمايش‌شده Crawled اتفاق افتاده است. پيمايشگر با يك روال منظم، مثلا يك يا دو بار در ماه به سايت مراجعه مي‌كند تا تغييرات موجود در آنرا بيابد. هر چيزي كه پيمايشگر بيابد به عنصر دوم يك موتور جستجو يعني فهرست انتقال پيدا مي كند. فهرست اغلب به كاتالوگي بزرگ اطلاق مي‌شود كه شامل ليستي از آنچه است كه پيمايشگر يافته است. مانند كتاب عظيمي كه فهرستي را از آنچه كه پيمايشگرها از صفحات وب يافته‌اند، شامل شده است. هرگاه سايتي دچار تغيير شود، اين فهرست نيز به روز خواهد شد.

از زماني كه تغييري در صفحه‌اي از سايت ايجاد شده تا هنگاميكه آن تغيير در فهرست موتور جستجو ثبت شود مدت زماني طول خواهد كشيد. پس ممكن است كه يك سايت پيمايش‌شده باشد اما فهرست شده نباشد. تا زماني كه اين فهرست‌بندي براي آن تغيير ثبت نشده باشد، نمي‌توان انتظار داشت كه در نتايج جستجو آن تغيير را ببينيم. نرم‌افزار موتور جستجو، سومين عنصر يك موتور جستجو است و به برنامه‌اي اطلاق مي‌شود كه بصورت هوشمندانه‌اي داده‌هاي موجود در فهرست را دسته‌بندي كرده و آنها را بر اساس اهميت طبقه‌بندي مي‌كند تا نتيجه جستجو با كلمه‌هاي درخواست شده هر چه بيشتر منطبق و مربوط باشد.

چگونه موتورهاي جستجو صفحات وب را رتبه‌بندي مي‌كنند؟

وقتي شما از موتورهاي جستجوي پيمايشي چيزي را براي جستجو درخواست مي‌نماييد، تقريبا بلافاصله اين جستجو از ميان ميليونها صفحه صورت گرفته و مرتب مي‌شود بطوريكه مربوطترين آنها نسبت به موضوع مورد درخواست شما رتبه بالاتري را احراز نمايد.

البته بايد در نظر داشته باشيد كه موتورهاي جستجو همواره نتايج درستي را به شما ارائه نخواهند داد و مسلما صفحات نامربوطي را هم در نتيجه جستجو دريافت مي‌كنيد و گاهي اوقات مجبور هستيد كه جستجوي دقيقتري را براي آنچه كه مي‌خواهيد انجام دهيد اما موتورهاي جستجو كار حيرت‌انگيز ديگري نيز انجام مي‌دهند.

موتورهای جستجو

فرض كنيد كه شما به يك كتابدار مراجعه مي كنيد و از وي درباره «سفر» كتابي مي خواهيد. او براي اينكه جواب درستي به شما بدهد و كتاب مفيدي را به شما ارائه نمايد با پرسيدن سوؤالاتي از شما و با استفاده از تجارب خود كتاب مورد نظرتان را به شما تحويل خواهد داد. موتورهاي جستجو همچنين توانايي ندارند اما به نوعي آنها را شبيه‌سازي مي‌كنند.

پس موتورهاي جستجوي پيمايشي چگونه به جواب مورد نظرتان از ميان ميليونها صفحه وب مي‌رسند؟ آنها يك مجموعه از قوانين را دارند كه الگوريتم ناميده مي‌شود. الگوريتمهاي مورد نظر براي هر موتور جستجويي خاص و تقريبا سري هستند اما به هر حال از قوانين زير پيروي مي‌كنند:

مكان و تكرار

يكي از قوانين اصلي در الگوريتمهاي رتبه‌بندي موقعيت و تعداد تكرار كلماتي است كه در صفحه مورد استفاده قرار گرفته‌اند كه ناميده مي‌شود (Location/Frequency بطور خلاصه روش مكان-تكرار (Methode

كتابدار مذكور را به خاطر مي‌آورِد لازم است كه او كتابهاي در رابطه با كلمه «سفر» را طبق درخواست شما بيابد. او در وحله اول احساس مي‌كند كه شما به دنبال كتابهايي هستيد كه در نامشان كلمه «سفر» را شامل شوند. موتورهاي جستجو هم دقيقا همان كار را انجام مي‌دهند. حاوي HTML موجود در كد Titleآنها هم صفحاتي را برايتان ليست مي‌كنند كه در برچسب كلمه «سفر» باشند.

موتورهاي جستجو همچنين به دنبال كلمه مورد نظر در بالاي صفحات و يا در ابتداي پاراگرافها هستند. آنها فرض مي‌كنند كه صفحاتي كه حاوي آن كلمه در بالاي خود و يا ابتداي پاراگرافها و عناوين باشند به نتيجه مورد نظر شما مربوطتر هستند.

عامل بزرگ و مهم ديگري است كه موتورهاي جستجو از طريق آن Frequencyتكرار يا صفحات مربوط را شناسايي مي‌نمايند. موتورهاي جستجو صفحات را تجزيه كرده و با توجه به تكرار كلمه‌اي در صفحه متوجه مي‌شوند كه آن كلمه نسبت به ديگر كلمات اهميت بيشتري در آن صفحه دارد و آن صفحه را در درجه بالاتري نسبت به صفحات ديگر قرار مي‌دهند.

موتورهای جستجو

عوامل خارج از صفحه

موتورهاي جستجوي پيمايشي اكنون تجربه فراواني در رابطه با وب مسترهايي دارند كه صفحات خود را براي كسب رتبه بهتر مرتبا بازنويسي مي‌كنند. بعضي از وب مسترهاي خبره حتي ممكن است به سمت روشهايي مانند مهندسي معكوس براي كشف چگونگي روشهاي مكان-تكرار بروند. به همين دليل، تمامي موتورهاي جستجوي معروف از روشهاي امتيازبندي «خارج از صفحه» استفاده مي‌كنند. عوامل خارج از صفحه عواملي هستند كه از تيررس وب‌مسترها خارجند و آنها نمي‌توانند در آن دخالت كنند و مساله مهم در آن تحليل ارتباطات و لينكهاست. بوسيله تجزيه صفحات، موتورهاي جستجو لينكها را بررسي كرده و از محبوبيت آنها مي‌فهمند كه آن صفحات مهم بوده و شايسته ترفيع رتبه هستند. بعلاوه تكنيكهاي پيشرفته به گونه‌اي است كه از ايجاد لينكهاي مصنوعي توسط وب‌مسترها براي فريب موتورهاي جستجو جلوگيري مي‌نمايد. علاوه بر آن موتورهاي جستجو بررسي مي‌كنند كه كدام صفحه توسط يك كاربر كه كلمه‌اي را جستجو كرده انتخاب مي‌شود و سپس با توجه به تعداد انتخابها، رتبه صفحه مورد نظر را تعيين كرده و مقام آنرا در نتيجه جستجو جابجا مي‌نمايند.

نتيجه‌گيري‌ها

اين پژوهش ٨ موتور كاوش اصلي را با دوبار جستجوي ٢٦ سؤال در هر كدام از آنها (٤ بار در اكسايت) ارزشيابي كرد. اطلاعات نشان داد كه موتورهاي كاوش انتخابي نمي‌توانند نتايج خوبي براي سؤالات مرجع واقعي ارائه دهند. اما در مورد سؤالات موضوعي ساختگي خوب عمل كردند. اين نكته نيز فهميده شد كه موتورهاي كاوش براي دو نوع سؤال بطور متفاوت عمل كردند: اينفوسيك در سؤالات موضوعي بهتر عمل كرد؛ در حاليكه اين تكست در سؤالات مرجع بهترين بود. از اين پژوهش فهميده شد كه موارد تكراري يك مشكل نمي‌تواند باشد. با تعريف متغير تنظيم براساس ميزان ارتباط در موتورهاي كاوش مي‌تواند ارزشيابي شود. با ارائه اطلاعاتي در مورد اين ٤ متغير اين پژوهش چندين جنبه از عملكرد موتورهاي كاوش را روشن كرد.

اين پژوهش بدون محدوديت نبود. اول، داده‌هاي آن لحظات ناپايدار را در اينترنت ثبت كرد. آنها

عكسهايي ارائه دادند از اينكه چگونه موتورهاي كاوش از آوريل تا ژوئن ۱۹۹٦ كار كردند. و اين تصاوير ممكن است كاملاً نهايي نباشد كه قبلاً بود زيرا اينترنت سريعاً در حال گسترش است. با اين وجود، مشابه مطالعه ليتون، اين پژوهش دريافت كه اينفوسيك يكي از بهترين موتورهاي كاوش است. اگر اين تحقيق تكرا شده و همان يافتهها كه بدست آمده. به هر حال يكي از تحقيقات قادر خواهد بود كه اطمينان بيشتري در مورد اين يافتهها بدهد. دوم، سؤالات مرجع از يك كتابخانهء دانشگاهي جمعآوري شده بود و سخت بود تعيين كردن اينكه آنها نمونهاي از سؤالات مرجع بودند. باز هم، تكرار اين پژوهش اعتبار يافتهها را افزايش ميداد. سوم، تعداد سؤالات آزمايش احتمالاً ميتوانست زياد باشد اگرچه اين پژوهش سؤالات بيشتري نسبت به اغلب پژوهشهاي ديگري بكار برد.

گذشته از محدوديتها، اين پژوهش متغيرهاي جديد براي ارزشيابي تنظيم براساس ارتباط توليد كرد و يك طرح تحقيقي براي مقايسه عملكرد موتورهاي كاوش براي دو نوع سؤال بكار برد، درك عملكرد موتورهاي كاوش را افزايش داد، توصيههايي در مورد اينكه چگونه طراحان سيستم ميتوانند سيستمهاي خود را بهبود بخشند ارائه داد و اشاره كرد كه چگونه كتابداران ميتوانند خودشان و مردم را براي جستجو در اينترنت آماده كنند.

جدول۱: ميانگين مانعيت و نقاط كور براي سؤالات مرجع در موتورهاي كاوش

موتورها	دقت	نقاط كور	درصد
آلتاويستا	05/2	21	50%
اكسايت	75/1	12	29%
اينفوسيك	95/1	19	45%
لايكاس	93/1	16	38%
ماژلان	33/1	27	64%
اپن تكست	93/2	15	36%
وب كرولر	10/1	24	57%
ورلد وايد وب	31/0	32	76%

			وارم

جدول ۲- ميانگين مانعيت و تعداد نقاط كور موتورهاي كاوش براي سؤالات موضوعي

موتورها	دقت	نقاط كور (فراواني)	دقت خاص (فراواني)
آلتاويستا	4/5	0	3
اكسايت	2/4	1	3
اينفوسيك	3/7	1	8
لايكاس	3/6	0	6
ماژلان	7/6	1	7
اپن تكست	5/6	2	6
وب كرولر	3/5	0	6
ورلد وايد وب وارم	2/3	5	2

1. دقت خاص به تعداد جستجوهايي كه بيش از ٥ گزينه مرتبط را بازيابي كردند اشاره دارد.

مهم ترين موتور هاي جستجو

www.google.com

http://www.yahoo.com/

موتورهای جستجو

http://www.37.com/

http://www.infoseek.com/

http://www.aliweb.com/

http://www.ask.com/

http://www.mamma.com/

http://www.altavista.com/

منابع:

www.irandoc.ac.ir/ETELA-ART/ 18/18_3_4_10.htm

http://www.tarighat-e.com/information/ShowArticle.asp?ID=341

پي نوشت ها:

موتور های جستجو

1. College & Research Libraries News.
2. Data collection
3. Precision
4. Duplicate
5. Most-relevant-item scove (MRI)
6. Releven cy-van king score.
7. Recell
8. Data analysis

موتور های جستجو

نویسنده: ابراهیم آرام

.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.
.

موتور های جستجو

این کتاب به صورت رایگان ارائه میشود.

موتور های جستجو