تجزیه و تحلیل موتورهای جست و جو در وب [1]

نوشته کارستن شلیکتینگ و اریک نیلسن

ترجمه مهدی نعمتی و حسن طالب زاده [2]

چکیده

این نوشته گزارشی است پیرامون چند موتور جست وجوی پراستفاده در وب جهانگستر (Lygos , Infoseek , Exite , Altavista) . پنج عضو هیأت علمی سیاهه ای از چهار تا شش کلید واژه برای یک موضوع خاص تحقیقاتی فراهم کردند . بعد از اجرای این کلید واژه ها در موتورهای جست وجو ، ده سایت بازیابی شده از طریق هر موتور جست وجو مورد ارزیابی قرار گرفت و نتایج از نظر کیفیت دسته بندی شد . تحلیل وارسی دو معیار برای هر موتور جست وجو در نظر می گیرد . اولین معیار (d') ، حساسیت موتور جست وجو را در پیدا کردن اطلاعات مفید مورد سنجش قرار می دهد . معیار دوم (Beta) ، میزان محدودیت یا آزادی موتور جست وجو را در تعیین این که چه سایت هایی در پاسخ به بازیابی در بر می گیرد مشخص می کند . نتایج ، نمایانگر عملکرد کاملا ضعیف تکنولوژی حاضر است و یک روش تحقیقاتی عینی را برای تجزیه و تحلیل پیشرفت های آینده فراهم می سازد .

---

۱ - این نوشته ترجمه متن زیر است : Carsten Schlichting & Erik Nilsen . Signal Delection Analysis of www Search Engines . Lewis & clark college portland , or 97219 USA +1 503 768 7657 carsten , nilson @ clark .

۲ - کارشناسان مرکز کامپیوتر ، اطلاعات و آمار دانشگاه فردوسی مشهد

مقدّمه

ارتباطات جهانی به تدریج با اهمیّت تر می شوند و میزان اطلاعات موجود پا به پای افزایش زیر ساخت ارتباطی پیش می رود که از دیدگاه نظری ، ظرفیت بالای نظام های حافظه الکترونیکی ، ما را قادر می سازد تا مقدار زیادی از اطلاعات را نگهداری کنیم .

تکنولوژی وب جهانگستر این اطلاعات را برای کل جامعه ارتباطی فراهم می سازد . به موازات انفجار اطلاعات ، مشکل بازیابی اطلاعات مرتبط و مناسب نیز وجود دارد . موتورهای جست وجو مشهورترین ابزارهایی هستند که برای جست وجوی منابع موجود در وب مورد استفاده قرار می گیرند . اخیراً تعداد بسیار زیادی از موتورهای جست وجو در دسترس قرار گرفته اند .

بازنگری نوشته ها در باره موتورهای جست وجو هم به صورت چاپی (TENNAT, 1996) و همچنین به صورت الکترونیکی پیوسته (Liu , 1996) به چشم می خورد . این بازنگری ها معمولاً به جنبه های فنی موتورهای جست وجو مانند حجم (اندازه) بانک اطلاعاتی ، تجربهٔ نویسندگان و اولویت های فردی می پردازند .

هر موتور جست وجو ، الگوریتم متفاوتی را برای نمایه کردن اطلاعات موجود در وب دنبال کرده ، نتایج را در ازای درخواست استفاده کننده نشان می دهد . برای بهره گیری مؤثرتر از وب لازم است که از موتور جست وجویی استفاده کنید که با حوزهٔ موضوع شما مناسبت بیشتری دارد . با این حال ، بسیاری از نوشته های بازنگری فاقد رتبه بندی هستند که بتواند در تصمیم گیری مربوط به گزینش یک موتور جست وجوی خاص کمک کند . برخی از رتبه بندیهای موجود فاقد مبنای علمی اند و تنها با اختصاص چند ستاره به موتورهای جست وجو تمایز بین موتورهای جست وجو را می نمایانند (ARENTS , 1996) . در سال ۱۹۹۵ وین شیپ (Winship) برای ارائه یک مقایسه عملکردی تلاش کرد . جدول ٤ از مقالهٔ او سه درخواست جست وجوی متفاوت و تعداد نتایج فراهم شده در شش بانک اطلاعاتی را نشان می دهد . وین شیپ به این نکته اشاره می کند که براساس تعداد نتایج به دست آمده نمی توان در مورد سودمندی موتورهای جست وجوی مختلف نتیجه گیری کرد زیرا پیوندهای تکراری و نامربوط موجب تحریف سنجش می شود . لیتُن (Leighton) در

سال ۱۹۹۵ این مسائل را در نظر گرفت و فقط تعداد پیوندهای مربوط را شمرد و از پیوندهای تکراری چشم پوشید . امّا او خود درخواست هایی برای جست وجو ارائه داد که به نتیجه ای سوگیرانه منجر شد .

با ظهور موتورهای جست وجوی بزرگ مانند Altavista که بیش از سی میلیون صفحهٔ خانگی را تحت پوشش دارد ، اندازه گیری تعداد موارد یافته شده دیگر یک اندازه گیری مؤثّر نیست . مسئلهٔ اصلی و مهم این است که کیفیت یافته ها مهمتر از کمیّت آنهاست . یک رهبرد برای غلبه بر این معضل اطلاعاتی ، استفاده از راهنمای وب است که صفحات کمتری دارد و این صفحات پس از طی نوعی مراحل آزمایشی کیفی به بانک اطلاعاتی وارد می شود . یکی از مشکلات این رهبرد ،آن است که امکان دارد مدیریت راهنمای وب و استفاده کنندگان از آن راهنما برای کیفیت استانداردهای متفاوتی در نظر داشته باشند . افزون بر این ، ممکن است استفاده کننده به دنبال برخی اطلاعات شخصی باشد که در چنین نمایه ای منظور نشده است . این امر سبب می شود که استفاده کننده به سراغ موتور جست وجوی جامعتری برود .

هدف از این تحقیق نشان دادن روشی است که بتوان از آن به طور عینی برای مقایسه موتورهای جست وجو با روش های کشف منابع هوشمند استفاده کرد . برای آن که یک چنین سنجش عملکرد ، اعتبار داشته باشد ، لازم است نوعی ارزشیابی از کیفیت نتایج به دست آمده از موتورهای جست وجو دراختیار باشد .

روش

از پنج عضو هیئت علمی دانشکده لوئیس و کلارک (Lowis and Klark) خواسته شد که درباره اطلاعات مشخّصی که میل دارند در وب پیدا کنند و هنوز در آن موارد دست به جست وجو نزده اند فکر کنند .

آنها اطلاعاتی را که به دنبال آن بودند در یک پاراگراف کوتاه توصیف کردند . علاوه بر این ، هر درخواست جست وجو را به صورت چهار تا شش کلید واژه فرمول بندی کردند . آن کلید واژه ها به ترتیب اهمیّت و ارتباط با موضوع جست وجو مرتب شدند . سپس ،

پژوهشگران از این کلید واژه ها برای جست وجو در چهار موتور جست وجوی (Excite , Infoseek , Lycos , Altavista) استفاده کردند .

موتورهای جست وجویی چون (Infoseek , Altavista) که جست وجوی عبارتی مثل " اندازه گیری رنگ ها " در آنها میسر بود ، مورد استفاده قرار گرفتند . ده مورد یافته های اول هر چهار موتور جست وجو در یک مدرک گردآوری شد . اگر یک موتور جست وجو کمتر از ده مورد یافته را بازیابی کرده بود ، این نتایج به مدرک اضافه می شد و جست وجو با حذف کلید واژه های کمتر مهم دوباره به اجرا در می آمد . از نتایج جدید ، فقط چند مورد اول از یافته ها مورد استفاده قرار گرفت تا تعداد آنها با موارد قبلا یافت شده به ده برسد . سپس آن نتایج توسط اعضای هیات علمی متقاضی مورد بررسی قرار گرفت و براساس میزان سودمند بودن و ربط یافته ها با جست وجو به آنها نمره داده شد . اعضای هیات علمی ابتدا در این مورد که یافته شده ها با جست وجوی آنها مربوط است یا نه تصمیم گرفتند . اگر مورد یافته شده مربوط بود ، میزان مفید بودن آن براساس یک مقیاس ، از یک تا هفت مشخص می شد . عدد یک کمترین میزان مفید بودن و عدد هفت بیشترین میزان مفید بودن را می نمایاند .

نتایج

ساده ترین گزارش یافته ها این بود که تعداد پیوندهای مربوط به دست آمده از هر موتور جست وجو شمارش شود . از مجموع دویست پیوند ارائه شده، با پنج جست وجو ، مجموعا تعداد پنجاه و چهار پیوند مربوط به موضوع پیدا شد.

لایکوز (Lycos) بیشترین تعداد پیوندهای مربوط را با نوزده مورد پیدا کرد ؛ اکسایت (Excite) با چهارده مورد ، اینفوسیک (Infoseek) با دوازده مورد وآلتا ویستا (Altavista) با نه مورد پیوند مربوط در ردیف های بعدی قرار داشتند . مشکل گزارش منحصربه تعداد پیوندهای مربوط ، آن است که در آن از تعداد کمی از داده ها استفاده می شود و از گسترهٔ اطلاعات گسترده تر در مورد جست وجو استفاده نمی شود . روش تحلیل و بررسی نشانه ها (Signal Detection Analysis = SDA) ، امکان توجه به جزئیات بیشتر را در مورد

عملکرد موتورهای جست وجو فراهم می آورد . این کار با ادغام اطلاعات بیشتر در یک چهارچوب یک پارچه انجام می گیرد .

نخستین قدم در روش SDA این است که پیوندهای بازیابی شده توسط هر یک از چهار موتور جست وجو را در یکی از چهار مقوله دسته بندی کنیم . شکل (۱) مقوله ها را نشان داده و توضیح می دهد . این مقوله بندی براساس قضاوت بلی / خیر در مورد مربوط بودن هر پیوند مبتنی است . این مرحله برای هر موتور جست وجوی ناموفق در پنج درخواست بطور جداگانه انجام شد .

قدم بعدی عبارت است از تعیین درجهٔ بازیافته ها و درجهٔ بازیافت کاذب در مورد هر موتور جست وجو درجه بازیافته ها با نسبت پیوندهای مفید (یافته های مناسب) ارائه شده بوسیلهٔ هر موتور جست وجو در ارتباط با تعداد کل پیوندهای مفید ارائه شده بوسیلهٔ تمام چهار موتور جست وجو (یافته ها + نایافته ها) تعریف می شود . به همین ترتیب ، نرخ (درجه) یافته های کاذب عبارت است از نسبت پیوندهای نامربوط ارائه شده بوسیلهٔ هر موتور جست وجو در ارتباط با تعداد کل پیوندهای نامربوط ارائه شده توسط همهٔ موتورهای جست وجو (یافته های کاذب + رد شده های درست) . در حالت ایده آل ، یک موتور جست وجو باید درجه بالایی از یافته ها و درجه پایینی از یافته های کاذب را به دست دهد . درجه های به دست آمده در شکل (۲) گزارش شده اند .

	Reported Links	Unreported Links
Good Links	**Hit** Relevant Site Found by target Search engine	**MISS** Relevant Site not Found by target Search engine but found by others
Bad Links	**False Alarm** Irrelevant Site Found By target Search engine	**Correct Rejection** Irrelevant Site not Found by target Search engine but found by others

شکل (۱) - دسته بندهای روش SDA برای پیوندهای موتور جست وجو

Hit Rates			
Lycos	Excite	Infoseek	Alta Vista
37.3 %	27.5 %	23.5 %	17.6 %

False Alarm Rates			
Lycos	Excite	Infoseek	Alta Vista
21.8 %	25.4 %	26.8 %	28.9 %

شکل (۲) – درجه های یافته ها و بازیافت های کاذب ترکیب شده برای هر پنج درخواست جست وجو

با استفاده از فرمول های استاندارد و جدول مبتنی بر درجات یافته های مفید و یافته های کاذب (Swets , 1964) تنظیم شده است ، تحلیل SDA دو نمره برای هر موتور جست وجو ارائه می دهد . یک نمره (d') حساسیت موتور جست وجو را دریافتن اطلاعات مفید اندازه گیری می کند . هر چه نمره (d') بیشتر باشد بهتر است . طیف معمول نمره (d') از صفر تا دو می باشد . نمرهٔ صفر به این مفهوم است که موتور جست وجو قادر به تمایز میان پیوندهای مفید و نامناسب از یکدیگر نیست .

نمره دیگر (Beta) میزان تحریف را در تصمیم گیری و این که تا چه میزان موتور جست وجو در گزارش کردن سایت ها محافظه کارانه یا آزادانه عمل می کند ، می نمایاند . هر چه مقدار Beta بیشتر باشد ، موتور جست وجو در گزارش کردن سایت ها محافظه کارانه تر عمل می کند . در این مفهوم ، رفتار محافظه کارانه یعنی تلاش برای در حداقل نگاه داشتن تعداد یافته های کاذب که در نتیجه برخی از یافته ها را از دست می دهد ، در حالی که آزادانه عمل کردن یعنی قبول درجه بالاتری از یافته های کاذب به جای گزارش کردن بالاترین درصد یافته ها از جنبه نظری . این مقیاس ها قابل تفکیک اند . نمره های به دست آمده در شکل (۳) نشان داده می شوند .

d' (measure of sensitivity)			
Lycos	Excite	Infoseek	Alta Vista
.48	.07	- .15	- .42
B (response bias)			
Lycos	Excite	Infoseek	Alta Vista
.40	.94	?	?

شکل (۳) – رکوردهای نمره ای SDA در مورد هر موتور جست وجو برای هر پنج درخواست ترکیب شده است

از چهار موتور جست وجو فقط لایکوز ، یک نمره قابل قبول (d') نشان داد . مقادیر منفی مشخصا در مورد اینفوسیک و آلتا ویستا مشکل آفرین است . این مقادیر نشان می دهند که عملکرد موتورهای جست وجو با این آزمایش آنچنان ضعیف اند که با مفروضات استاندارد SDA تناسب ندارند . یک فرض اصلی این است که توزیع سایت های دارای پیوندهای مربوط (یافته ها + نایافته ها) احتمالا بیشتر توسط مکانیسم جست وجو نمایه سازی می شود تا سایت های دارای پیوندهای نامربوط (یافته های کاذب + رد شده های درست) این نقض فرضیه موجب پیچیده تر شدن تفسیر نمره Beta می شود . لایکوز در مقایسه با اکسایت تقریبا سطح معیار آزادانه تری دارد . با این حال ، محاسبات Beta ، زمانی که نتایج (d') منفی است قابل تفسیر نیست . بنابراین نمی توان میزان سوگیری پاسخ ها را در اینفوسیک یا آلتاویستا تعیین کرد .

یکی دیگر از نتایج جالب این پژوهش این است که میان پیوندهای پیدا شده توسط چهار موتور جست وجو تقریبا هیچ هم پوشانی وجود ندارد . از مجموع بیش از ۲۰۰ پیوند ، فقط ۵ مورد در دو موتور جست وجو هم پوشانی داشت . با در نظر گرفتن جزئیات درخواست ها ، این یک نتیجهٔ تعجب آور است .

بحث

آشکارترین نتیجه این است که عملکرد موتورهای جست وجو از وضعیت ایده آل بسیار دور است . هیچ یک از موتورهای جست وجو به سطح قابل قبولی از عملکرد در مورد این درخواست های کاملا مشخص و علمی نرسیده اند . کاربرد موفقیت آمیز روش SDA مستلزم بهبود فن آوری جست وجو یا تغییر در توجه به این بررسی است .

تحقیقات آینده از روش SDA برای اندازه گیری میزان اثر بخشی راهبردهای مختلف جست وجو در یک موتور جست وجو استفاده خواهد کرد . برای مثال ، در مقابل هم قرار دادن عملگرهای منطقی OR و AND در ساختار درخواست های جست وجو . سایر احتمالات عبارت است از مقایسهٔ جست وجوی " مفهوم " ، با جست وجوی کلید واژه ای با استفاده از اکسایت و تغییر تعداد کلید واژه های استفاده شده در یک درخواست جست وجو . این تحقیق یک آزمایش تجربی از رهنمودهای ارائه شده به وسیلهٔ بسیاری از موتورهای جست وجو برای ساخت و پرداخت یک جست وجوی موفق بدست خواهد داد .

این بررسی فقط اولین ده سایت ارائه شده از هر موتور جست وجو را مورد استفاده قرار داد زیرا فرض شده بود که استفاده کنندگان فراتر از یک صفحه اطلاعات را مورد جست وجو قرار نمی دهند . مطالعه و پیگیر ، این فرضیه را بصورت تجربی مورد آزمایش قرار خواهد داد . با استفاده از یک رایانه ردیاب عملا خواهیم دید که هنگام نشان دادن نتایج جست وجو ، استفاده کنندگان به چند سایت نگاه خواهند کرد ، همچنین آماری تهیه شود که مشخص کند آنها به کجای صفحه رایانه چشم خواهند دوخت .

درون داد ، این مقاله به علم شناسایی یک روش عینی برای ارزیابی میزان اثر بخشی تکنولوژی های موجود و آینده در مورد یافتن منابع است . هدف این تحقیق ، این نیست که نشان دهد کدام یک از موتورهای جست وجو برتر است .

تا زمانی که این مقاله ارائه شود ممکن است تکنولوژی برتری در دسترس باشد . با تکنولوژی جست وجوی بهتر ، SDA این توانمندی بالقوه را خواهد داشت که از گذاشتن مقطعی پیرامون عملکردهای شخصی ، شمارش ساده تعداد پیوندهای بازیابی شده ، یا تعداد ستاره ها یا نشانه هایی که یک فرد بررسی کننده به یک سایت می دهد ، فراتر رود .