خزنده و ساختواره وب

شعله ارسطوپور[1]

چکیده

وب به عنوان بستر فعالیت موتورهای جستجو، ساختاری نمـوداری دارد. ایـن سـاختار حرکـت خزنـده‌هـا در موتورهای جستجو را به روشهایی منطبق بر خود محدود می‌سازد. مقاله حاضر، به بررسی تأثیر ساختار وب بر چگـونگی حرکت خزنده‌ها و فعالیت نمایه سازها در موتورهای جستجو می پردازد. پس از بحثی مقـدماتی در بـاب نمودارهـای جهت‌دار و فرایند کار خزنده، عمده‌ترین روشهای حرکت خزنده در سطح وب شامل حرکت «عمق ـ شروع»، «توزیع ـ شروع» و «بهترین ـ شروع» مطرح شده و سپس واحد سازه‌یابی و چگونگی تشکیل درختهای سازه‌یابی از قالب HTML مورد بررسی قرار خواهد گرفت.

***کلیدواژه‌ها*: ساختار نموداری وب، خزنده‌ها، حرکت عمق ـ شروع، حرکت توزیع ـ شروع، حرکت بهتـرین ـ شروع، نمایه‌سازی وب**

مقدمه

وب، مجموعه‌ای عظیم از مدارک است که هر یک برای پاسـخگویی بـه یـک نیـاز بـالقوه ایجـاد و در بستر اینترنت منتشر گردیده است. حجم عظیم صفحات و اطلاعـات موجـود در وب و لـزوم وجـود ابزارهـایی جهت سازماندهیِ دست کم گوشه‌ای از این مجموعه تاکنون به حد کفایت مـورد بحـث قـرار گرفتـه اسـت و اینکه موتورهای جستجو با تمام کاستیهای خود از عمده‌ترین ابزارهای دسترسی به محتوای مدارک پیش گفته هستند، اصلی پذیرفته شده میان بسیاری از کاربران اینترنتی است. وب، میلیاردها صفحه الکترونیکی را از طریق شبکه ای از اتصالها با یکدیگر مرتبط می سازد و از آنجا کـه سـاختار آن هـیچ محـدودیتی بـر چگـونگی نـشر صفحات، قالب انتشار، و یـا چگـونگی برقـراری اتـصالها و تعـداد صـفحات وب اعمـال نمـی‌نمایـد، گـسترش چشمگیری یافته است. اما این مسئله هرگز بدین معنا نیست که وب بدون سـاختار اسـت. بـسیاری از پژوهـشها بیانگر وجود روابطی مشخص و منطقی در ساختار وب هستند.

(Barabasi & Albert, 1999; Kleinberg etal., 1999; Albert etal., 1999).

هر یک از این تحقیقات به نوبه خود به نتایج متفاوتی رسیده اند. فصل مشترک آنها مـسئله ناهمـاهنگی توزیع صفحات وب است بـه گونـه‌ای کـه آلبـرت و دیگـران[2] (1999) از وب بـه عنـوان یـک شـبکه جهـانی کوچک یاد می‌کنند در حالی که برادر و همکاران[3] (2000) توزیع وب را پاپیونی شکل توصیف می‌کنند. بر

---

۱. دانشجوی کارشناسی ارشد کتابداری و اطلاع‌رسانی دانشگاه فردوسی مشهد

1. Albert etal.
2. Broder etal.

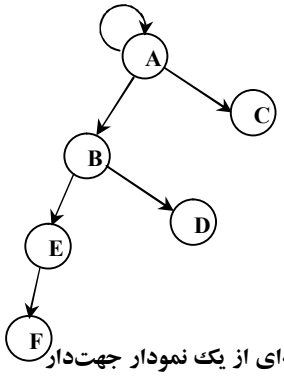
این اساس، وب از یک هسته که بیشترین اتصالها به آن برقرار شده و دو بال که با توجه به جهت اتصالها، توزیع متفاوتی از صفحات را در بر می‌گیرد، تشکیل شده است. کلیه پژوهش‌های انجام شده و تمام نتایج یک مسئله را اساس کار خود قرار داده اند و آن نموداری بودن ساختار وب[4] است. با توجه به دستاوردهای پژوهشها، ساختار وب را به یقین می‌توان به صورت یک نمودار عظیم ترسیم نمود؛ نموداری که در برگیرندهٔ گره‌ها و اتصالهای متعدد است. در این نمودار، صفحات وب همان گره‌ها و اتصالها همان لینکهای برقرار شده میان صفحات مختلف است
(Albert et al., 1999; Thelwall, 2002; Evans& Walker, 2004; Yu & Meng, 2004; Coathy, 2004) .

بنابراین وب نه تنها بدون ساختار نیست بلکه می‌توان آن را به صورت یک ساختواره نموداری نیز ترسیم نمود. این ساختواره خواسته یا ناخواسته بر چگونگی گردآوری صفحات و نمایه‌سازی آنها توسط موتورهای جستجو تأثیرگذار بوده و در نهایت نتایج جستجو و میزان ربط آنها را رقم می‌زند.

نمودارهای جهت‌دار[5]

نمودار در واقع مجموعه‌ای از گره‌ها و خطوط است که آن را به صورت ریاضیِ G (V, E) نشان می‌دهند. هر E حتماً دو گره را به یکدیگر متصل می‌کند. نمودار جهت‌دار، نموداری است که بتوان در آن جهت حرکت از هر گره به گره دیگر را به راحتی تشخیص داد.

در یک نمودار جهت‌دار، چنانچه تعداد اتصالها را با i نمایش دهیم، برای هر گره که به عنوان گره اصلی در نظر گرفته شود، رابطه $i \in [1,n)$ برقرار خواهد بود. در این شرایط، هر یک از اتصالها تنها به طرف یک گره حرکت خواهد کرد. از دیگر ویژگیهای عمده یک نمودار جهت‌دار آن است که هیچ کدام از گره‌ها منفرد و بدون اتصال نخواهد بود. (تصویر 1).



تصویر1. نمونه‌ای از یک نمودار جهت‌دار

پیشتر بیان شد که هر صفحه از وب دارای تعدادی اتصال است. این اتصالها یا از صفحه هسته[6] به سایر صفحات و یا از سایر صفحات به صفحه موردنظر برقرار شده است. بنابراین، تمام ویژگیهای پیش گفته، در باب صفحات وب و لینکهای آنها نیز صدق می‌کند. با توجه به جهت برقراری لینکها، می‌توان وب را نیز به صورت یک نمودار جهت‌دار ترسیم نمود.

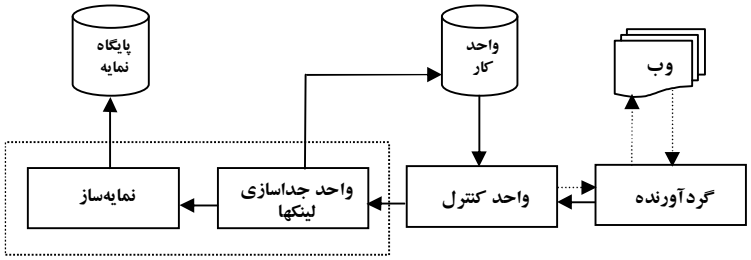
---

3. Web's Graph Structure
1. Directed Graph
2. Seed Page

خزنده

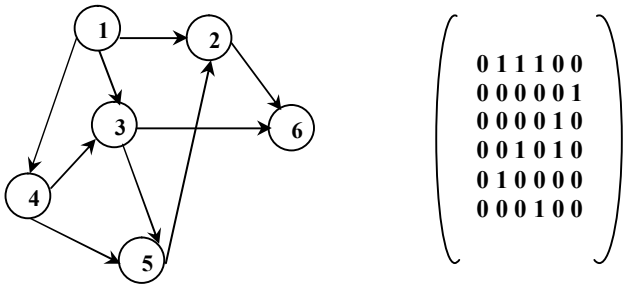
تقریباً تمام خزنده‌ها در موتورهـای جـستجو دارای چهـار بخـش گردآورنـده[7]، واحـد کنتـرل[8]، واحـد سازه‌یابی[9] و واحد کار[10] هستند. گردآورنده تحت نظارت واحد کنترل، به صفحات هسته (گره‌هـای مختلـف) رفته و مدارک را به واحد جداسازی لینکها می فرستد. پس از جداسازی، لینکهای مناسب به واحد کار ارسـال شده و در فهرست دستور کار بعدی گردآورنده قرار می‌گیرند. واحد سازه‌یابی در واقع از دو بخش جداسـازی لینکها[11] و نمایه‌سازی تشکیل شده است.

آنچه نهایتاً در پایگاه ذخیره می‌شود در واقع حاصل فرایند نمایه‌سازی است که تحت قالب تعریف شده در الگوریتم موتور جستجو به صورت واژگان و عبارات مختلف در آمده است (Cothey, 2004). تـصویر 2 ساختار یک خزنده را نشان می دهد.



تصویر 2. ساختار خزنده و چهار جزء اصلی آن

پیشتر اشاره شد که ساختار وب همچون نموداری جهت‌دار است، لذا خزنده نیز برای حرکت روی این ساختواره، چاره‌ای جز تبعیت از ویژگیها و جهت‌های از پیش تعیین شده ندارد. از نظر جبـری، حرکـت روی نمودارهای جهت‌دار را می توان به صورت ماتریسهای حرکتی[12] ساده کرد. در موتورهای جستجو نیز، با توجـه به شباهت ساختاری، استفاده از قواعد ریاضی حاکم بـر ایـن ماتریـسها، تحلیـل و مقایـسه را آسـان مـی‌نمایـد. (تصویر3)



تصویر3. ساختواره نموداری وب و ماتریس حرکتی متناظر با جهت‌های مشخص شده

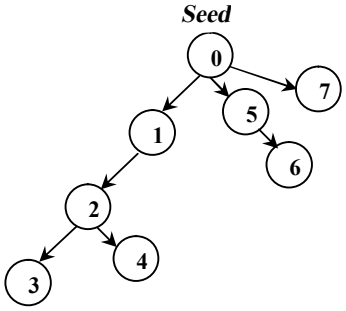
به طور کلی، سه روش برای حرکت خزنده در شبکه لینکهای وب وجود دارد. این سه روش عبارتند از حرکت عمق ـ شروع[13]، توزیع ـ شروع[14]، و بهترین ـ شروع[15].

---

1. Fetcher
2. Controller
3. Parsing Unit
4. Workload Unit
5. Link Extracting
6. Traversal Matrix
1. Depth - First
2. Breadth - First

حرکت عمق ـ شروع

در این حرکت، واحد کنترل خزنده یک صفحه را به عنوان صفحه هسته برای گردآورنده مشخص می‌سازد. پس از جداسازی لینکها، واحد کنترل یکی از لینکهای خارجی صفحه را انتخاب و گره مقصد را به گردآورنده معرفی می‌کند. این فرایند تا زمانی که برای واحد کنترل تعریف شده باشد، ادامه پیدا می‌کند. به عنوان نمونه، ترتیب حرکت گردآورنده به صفحه‌های مختلف با الگوریتم عمق ـ شروع، مانند تصویر 4 خواهد بود.

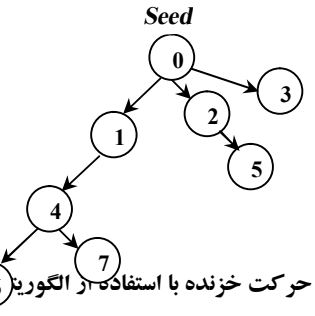


تصویر4. صفحه هسته و ترتیب حرکت خزنده با استفاده از الگوریتم عمق ـ شروع

از آنجا که تقریباً تمام صفحه‌های وب لینکهایی به سایر صفحات برقرار می‌کنند، چنانچه سطح عمق برای واحد کنترل تعریف شده نباشد، حرکت به عمق آن قدر ادامه خواهد یافت که به مرور زمان، عملاً پایگاه نمایه موتور جستجو را از مطالب بی کیفیت خواهد انباشت. به همین دلیل، در بیشتر موتورهای جستجو، سطح عمق برای واحد کنترل تعریف می‌شود. در تصویر4، چنانچه سطح عمق تعریف شده 2 باشد، ترتیب حرکت گردآورنده 1، 2، 5، 6و7 بوده و عملاً صفحه‌های 3 و 4 مورد بررسی قرار نخواهند گرفت.

حرکت توزیع ـ شروع

در این حرکت، واحد کنترل پس از تعیین صفحه هسته، کلیه گره‌های هم عمق با یکدیگر را تعیین و به ترتیب به گردآورنده معرفی می‌کند. پس از رجوع به کلیه صفحات مشخص شده در آن سطح، واحد کنترل سطح دوم را مورد بررسی قرار می دهد. به عنوان نمونه، ترتیب حرکت گردآورنده تحت نظارت واحد کنترل و با استفاده از الگوریتم توزیع ـ شروع در صفحه های مختلف مانند تصویر شماره 5 خواهد بود.



تصویر 5. صفحه هسته و ترتیب حرکت خزنده با استفاده از الگوریتم توزیع ـ شروع

---

3. Best - First

روش حرکت توزیع ـ شروع مورد علاقه بسیاری از طراحان برنامه‌های خزنده در موتورهای جستجوست، زیرا طراحی و اجرای آن به صورت رایانه‌ای بسیار ساده‌تر از روش حرکت عمق ـ شروع بوده و در صورت تعیین سیاست دقیق، به لحاظ محدود بودن دامنه لینکهای هر صفحه به عنوان صفحه هسته، حجم پایگاه موتور جستجو بیهوده افزایش نخواهد یافت(Chakrabarti et al., 2002).

حال، چنانچه صفحه هسته به یک مطلب خاص بپردازد، با توجه به آنکه گردآورنده تمام لینکهای موجود در صفحه و یا صفحات بعد را دنبال نمی‌کند، حرکت خزنده تأثیر بسیار زیادی بر نمایه‌سازی و در نهایت بازیابی اطلاعات خواهد داشت.

در حرکت عمق ـ شروع، با انتخاب هر لینک و رفتن به صفحه بعدی و ادامه این کار، یک مطلب خاص (حوزه موضوعی مربوط به سطح عمق اول حرکت) به صورت اختصاصی دنبال شده و از آنجا که گرایش واحد کنترل نسبت به حرکت عمقی گردآورنده بیشتر از حرکت در سطح است، در نهایت صفحاتی که برای نمایه‌ساز فرستاده می شوند به احتمال، اغلب حول یک مطلب یا موضوع خواهند بود. در حالی که در حرکت توزیع ـ شروع گرایش واحد کنترل به حرکت در سطح است و لذا گردآورنده ابتدا به گره‌های تعیین شده در سطح سرکشی خواهد کرد. در چنین شرایطی صفحاتی برای واحد نمایه‌ساز ارسال می شوند که در کل دیدی عامتر دارند (مانند آنچه در صفحه هسته آمده است). این مسئله، ناشی از آن است که معمولاً لینکهایی که از هر صفحه برقرار می شوند، به بخشی از مطلب مطرح شده در صفحه هسته مربوط می‌شوند.

بنابراین، با حرکت از سطح به عمق، دید نمایه‌ساز خواه نا خواه جزء نگر بوده و با حرکت در سطح دید نمایه‌سازی در حول یک مطلب با گستره‌ای وسیع‌تر و جامع‌تر متمرکز خواهد بود (Herrman, 2003).

با توجه به مطالعات دهه 1990، استفاده از هر یک از این روشها برای حرکت روی ساختواره وب با توجه به حجم عظیم صفحات و مطالب هر یک، در نهایت تفاوت چندانی در بازیابی بهتر موتورهای جستجوی مختلف در سطح وب نداشت.

تصمیم‌گیری در باب انتخاب هر یک از این دو روش و تردیدهای موجود، روش سومی را پیش پای طراحان و برنامه‌نویسان قرار داد و آن، حرکت «بهترین ـ شروع» بود.

حرکت بهترین ـ شروع

بهترین در حوزه حرکت خزنده روی ساختواره وب، در واقع معانی متفاوتی دارد. الگوریتمهای مختلفی برای حرکت بهترین ـ شروع وجود دارند که بر اساس فرمول محاسبه بهترین صفحه بعدی، اسامی متفاوت دارند. از این بین می‌توان به خزنده متمرکز[16]، جستجوی کوسه‌ای[17]، عنکبوتهای اطلاعاتی[18] و... اشاره کرد.

در ساده‌ترین حالت، از سیاستهای رتبه‌بندی همچون "رتبه‌بندی صفحات"[19] به عنوان معیار بهترین بودن استفاده می‌شود. در این روش واحد کنترل با توجه به رتبه هر صفحه میان سایر صفحات، گردآورنده را به صفحه بعدی می فرستد.

---

1. Focused Crawler
2. Shark Search
3. Info Spiders
4. Page Rank

حرکت بهترین ـ شروع، بر این اصل مبتنی است که پدیدآورنده هر صفحه زمانی از صفحه خود (A) به صفحه دیگری (B) لینک برقرار می‌کند که B از نظر پدیدآور A ارزشمند باشد. در بحث رتبه‌بندی صفحات چنین عملی برای B یک امتیاز مثبت محسوب می‌شود. بنابراین، در ساختار خزنده‌ها الگوریتم رتبه‌بندی صفحات در واقع برنامه‌ای است که اهمیت نسبی هر صفحه را بر اساس لینکهای برقرار شده به آن مشخص می‌سازد. بر این پایه در خزنده‌هایی که برای حرکت خود از روش بهترین ـ شروع استفاده می‌کنند در واقع از سه اصل پیروی می‌شود:

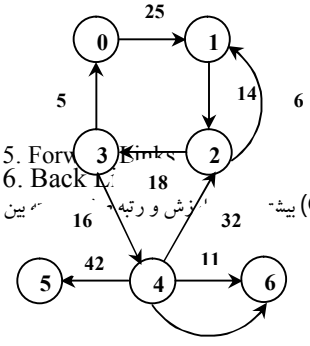
- صفحاتی که لینکهای بیشتری به آنها برقرار شده است، اهمیت بیشتری دارند. تعداد بیشتر لینکها به نوعی نشانگر شهرت و یا محبوبیت صفحه مذکور در سطح وب است.
- چنانچه این لینکها از صفحات معتبرتری برقرار شده باشند، اعتبار صفحه مورد مطالعه افزایش خواهد یافت.
- از طرفی، هرچه تعداد لینکهایی که از صفحه مورد مطالعه به سایر صفحات برقرار می‌شود بیشتر باشد، ارزش آن صفحات کمتر خواهد بود.

بر این اساس، در اکثر خزنده‌ها چنانچه U صفحه اصلی، F_u نشانگر صفحاتی که از U به آنها لینک برقرار شده[20] و B_u نشانگر صفحاتی باشد که به U لینک برقرار کرده‌اند[21] رتبه صفحه از طریق فرمول حاضر و یا فرمولهایی با عناصر اصلی مشابه، قابل سنجش خواهد بود.

$$R(U) = \sum_{V \in Bu} \frac{R(V)}{|F_u|}$$

در این فرمول، علامت جمع به معنای وجود رابطه مثبت میان تعداد لینکهای برقرار شده به U و رتبه U است. R(V) یا رتبه صفحات برقرار کننده لینک به U نیز چون در صورت کسر قرار گرفته اند، رابطه مثبت با رتبه U دارند، در حالی که وجود $|F_U|$ در مخرج، نشانگر وجود رابطه معکوس رتبه U و تعداد لینکهایی است که به صفحات دیگر بر قرار کرده است[22].

استفاده از نتیجه این فرمول در گردآوری صفحات ارزشمندتر، مؤثر خواهد بود. با کمک این روش و استفاده از رتبه‌بندی محتوای مدارک در واحد نمایه‌سازی به احتمال پاسخهای بهتری به نیاز کاربر موتور جستجو داده خواهد شد. با رتبه‌بندی صفحات، در واقع رتبه لینکهای منتهی به آنها مشخص شده و خود به خود ترتیب حرکت گردآورنده بر روی ساختواره نموداری عظیم وب روشن می گردد. تصویر 6 نشانگر این ترتیب بر اساس رتبه‌های مشخص شده در ماتریس حرکتی خواهد بود.



$$\begin{pmatrix} 0 & 25.0 & \infty & \infty & \infty & \infty & \infty \\ \infty & 0 & 10.0 & 14.0 & \infty & \infty & \infty \\ 5.0 & \infty & 0 & \infty & \infty & 16.0 & \infty \\ \infty & 6.0 & 18.0 & 0 & \infty & \infty & \infty \\ \infty & \infty & \infty & \infty & 0 & \infty & \infty \\ \infty & \infty & \infty & 32.0 & 42.0 & 0 & 14.0 \\ \infty & \infty & \infty & \infty & \infty & 11.0 & 0 \end{pmatrix}$$

5. For
6. Back

ـه (Child pages) بیش-
ـزش و رتبه ـ بین

۱. این مسئله تابع این اصل است که
این صفحات تقسیم می شود (Yu & Meng, 2004).

تصویر6. لینکها و ارزش عددی هر یک و نمایش آنها در قالب ماتریس حرکتی

این مسئله در نظر کاملاً منطقی است، اما ناجورک و وینر[23] (2001) در عمل ثابت می‌کنند که با توجه به هزینه بالا و زمان بر بودن فرایند رتبه‌بندی صفحات، استفاده از روش توزیع - شروع به نسبت توجیه پذیرتر می‌نماید. آنها برمبنای دستاوردهای خود بیان می‌دارند که با توجه به امکانات فعلی فناوری، روش رتبه‌بندی صفحات بسیار هزینه بر بوده، زمان زیادی را می‌طلبد و از طرف دیگر با توجه به عدم ثبات رتبه صفحات در طول زمان بایگانی نگهداری رتبه‌های صفحات را به سرعت باید روزآمدسازی نمود. این در حالی است که توجه به حرکت عمق ـ شروع به عنوان یک گزینه مطرح، کمتر صورت می‌پذیرد.

واحد سازه‌یابی

واحد نمایه ساز موتورهای جستجو باید صفحات حاوی اطلاعات را از گردآورنده دریافت کند و عبارات و واژگان آنها را استخراج و در پایگاه خود ذخیره سازی نمایند. بنابراین، چنانچه هر صفحه در مقام خود یک واحد کلی باشد، نمایه ساز آن را به اجزای کوچکتر از قبیل واژه و یا عبارت تبدیل کرده و در پایگاه خود ذخیره می‌سازد. نرم‌افزاری که توانایی انجام این عمل را داشته باشد، «سازه یاب» نام دارد. در فرایند سازه‌یابی، *اولین کار تشخیص زبان رشته نشانه‌های ورودی*[24] *است*. پس از آن، بر اساس دستور آن زبان خاص، سازه‌یاب به تعیین ساختار ترکیبی آن رشته می‌پردازد[25].

با این توصیف برنامه سازه‌یاب در ساختار یک موتور جستجو کار جداسازی و یکسان‌سازی آدرسهای اینترنتی موجود در مدرک، نگهداری فهرست واژگان غیر مجاز و تهیه درخت سازه‌یابی را انجام می دهد. از آنجا که سازه‌یاب براساس دستور زبان از قبل تعریف شده به هدف دستیابی به محتوای مشخصی عمل می‌کند، تقسیم‌بندی واژگان استخراج شده و وزن دهی به آنها کار ساده‌ای خواهد بود(Fischer, 2005)

زبان HTML زبان غالب در سطح وب به شمار می‌آید، لذا کلیه موتورهای جستجو دارای نرم افزارهای سازه‌یابی سازگار با HTML برای زبانهای مختلف هستند. به واسطه این نرم افزارها، برچسبهای HTML و ارزش آنها به سرعت شناسایی می شوند.

برای جداسازی و یکسان نمودن آدرسهای موجود در یک صفحه، از سازه‌یابها به منظور شناسایی برچسبهای مختلف و ارزش آنها استفاده می‌شود. این کار معمولاً به منظور کمک به واحد کنترل جهت هدایت گردآورنده انجام می‌شود. اما کار معرفی آدرسها به گردآورنده، اغلب مشکل‌تر از این است. گاه لازم است بسیاری از آدرسهای ذکر شده در صفحه یکدست و تصحیح شوند. به منظور یکسان سازی آدرسهای اینترنتی

---

1. Najork & Wiener
2. Input Symbols' String
۳. پاکروان، امیرحسین (۱۳۷۶)، فرهنگ کامپیوتر یادواره "انگلیسی - فارسی"، (تهران: یادواره اسدی؛ فرهنگستان یادواره).

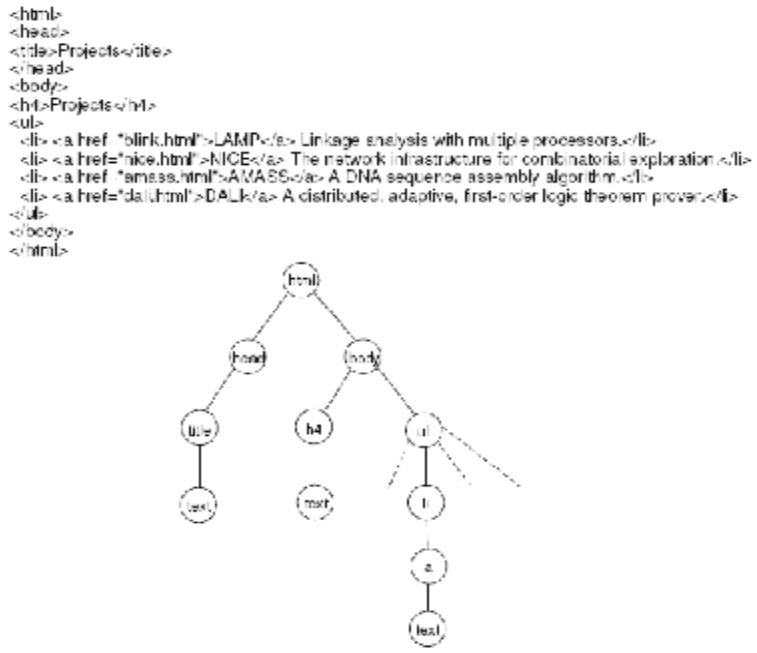
دستورالعملهایی برای تبدیل حروف بزرگ آدرسها به حروف کوچک، برداشتن انشعابهای اضافی از دنباله آدرس، تصحیح و یا تکمیل برخی از آدرسها و ...، برای سازه یابهای مختلف تعریف می‌شود (Pant, Sirinivan & Meczer, 2004).

سازه‌یابهای مختلف ممکن است سیاهه واژگان غیر مجاز متفاوتی داشته باشند و یا اصولاً فاقد ویژگی حذف واژگان بدون بار معنایی در طول فرایند نمایه‌سازی باشند. در سطوح بالاتر، برخی از سازه‌یابها با توجه به دستور زبان از پیش تعریف شده، توانایی تشخیص ریشه کلمات و ذخیره کلیه واژگان هم ریشه را در یک محل دارند.

در نهایت، وظیفه هر سازه‌یاب ایجاد درخت سازه‌یابی[26] است. در این مرحله، واحد سازه‌یابی آدرس و یا واژه موجود در صفحه را با کمک محتوا و محل برچسب ارزیابی کرده، درختواره‌ای از ساختار صفحه تشکیل می‌دهد. نمونه‌ای از این درخت و قالب HTML متناظر با آن، در تصویر 7 نشان داده شده است.

گره اول در این نمودار نشانگر قالب مدرک، گره‌های میانی برچسبهای مختلف مدرک و آخرین گره‌ها نماینده محتوای میان دو برچسب ابتدایی و انتهایی است.

تشکیل این درخت، کار وزن‌دهی به هر متن و واژه و عبارت استخراج شده از آن را ساده می‌نماید. رتبه متون در قسمتهای مختلف صفحه با توجه به الگوریتم رتبه‌بندی خاص هر موتور متفاوت است. واژگان وزن‌دهی شده را می‌توان به راحتی در قالب یک مقیاس عددی ریخته و برای الگوریتم رتبه‌بندی موتور جستجو، امکان سنجش و مقایسه سؤال کاربر و واژگان موجود پایگاه را فراهم آورد.



تصویر 7. درخت سازه‌یابی و قالب HTML متناظر با آن

جمع‌بندی

با وجود ماهیت متغیر وب، باز هم ساختار وب بر نحوهٔ سازماندهی آن تأثیرگذار خواهد بود. با توجه به اینکه هم اکنون موتورهای جستجو از مهمترین سازمان‌دهندگان به شمار می‌روند و از طرفی با در نظرگرفتن اینکه ساختار وب روشهای متفاوت و نه متعدد گردآوری اطلاعات را به طراحان نرم افزارهای خزنده دیکته

1. Parse Tree

می‌کند، می‌توان بیان داشت که ساختواره جهت‌دار وب بی‌تأثیر بر بازیابی‌های مفید و یا بی‌حاصل تأثیر خواهد داشت. آنچه در این زمان مورد توجه بیشتر محققان حوزه قرار گرفته، چگونگی بهینه‌سازی استفاده از امکاناتی است که وب در اختیار طراحان قرار می دهد. تصمیم‌گیری در باب انتخاب شیوه حرکت خزنده ـ اعم از حرکت به عمق یا حرکت در سطح و یا انتخاب هر صفحه بسته به کیفیت آن ـ یکی از مباحث مورد توجه علاقه‌مندان به این حوزه است. مطالعه در باب بازدهی هر روش در طول زمان و یا امکان‌سنجی استفاده از یک روش در حال حاضر از مطالعات مطرح در این حوزه به شمار می‌آید.

این در حالی است که از زاویه‌ای دیگر، اعمال از پیش تعریف شده برای هر سازه‌یاب ـ چه کوتاه و مختصر (سازه‌یاب‌های ساده) و چه پیچیده (سازه‌یابهای سطح بالا) ـ نیز به بهینه‌سازی استفاده از امکانات وب توجه می‌کند. با تعریف جزئیات بیشتر، نمایه‌سازی دقیق‌تر شده و در نهایت بازیابی بهتری حاصل خواهد شد.

مدارک به واسطهٔ شیوه حرکت در سطح وب گردآوری شده اند و نمایه‌سازی روی ساختار ساختمند وب به اجرا در آمده است؛ ساختاری که حتی زبان نگارش آن را می‌توان در قالب نمودار ترسیم نمود. الگوریتمهای مختلف رتبه‌بندی بر اساس این ساختار و اجزای آن کار خود را انجام می‌دهند، پس ساختار وب به صورتی غیر مستقیم اما با قدرتی بسیار بر آنچه بازیابی می‌شود تأثیر خواهد داشت.

منابع

- Albert, R., Jeoung, H.& Barabasi, A. (1999). "The Diameter of the Wold Wide Web". Nature . Vol. 401, P. 130 Available online: www10.org/cdrom/papers/208 [Accessed on Oct. 2005]

- Barabasi A.L. & Albert, R. (1999). "Emergence of Scaling in random Networks". Science. Vol. 286, P 509 - 512. Available online: http://www.nd.edu/~networks [Accessed on Oct. 2005]

- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, s., Stata, R., (2000). "Graph Structure in the Web". Computer Networks. Vol. 33, P. 309 - 320. Available online: http://www9.org/w9cdrom/160/160.html [Accessed on Oct. 2005]

- Chakrabarti, s. , Joshi, M.M., Punera, K. & Pennock, D.M. (2002). "The Structure of Broad Topics On the Web". Proceedings of the 11th World Wide Web Conference (p.508 – 510). Honolulu, Hawaii, May 7- 11 . New York: ACM. Available online: http://http.cs.berkeley.edu/~soumen/doc/www2002t/p338-chakrabarti.pdf [Accessed on Oct. 2005]

- Cothey, Viv (2004). "Web Crawling reliability". Journal of the American Society for Information Science and Technology. 55(14). P. 1228 – 1238.

- Evans, Michael P. & Walker, Andrew (2004). "Using The Web Graph to Influence Application Behavior". Internet Research. 14(5). P. 372 – 378.

- Fischer, Hendrik (2005). Decisions To Go: An Intelligent Mobile Decision Support System[Dissertation]. Georgia: The University of Georgia. Available online: http://graduate.gradsch.uga. edu/etdarchive/summer2005/fischer-hendrik-200508-ms.pdf [Accessed on Oct. 2005]

- Herrmann, Frank (2003). Web search engines. Available online: http://graduate.gradsch.uga.edu/etdarchive/summer2005/ fischer-hendrik-200508-ms.pdf [Accessed on Oct. 2005]

- Kleinberg, J., Kumar, R., Raghava, P., Rajagopalan, S., & Tomkins, A. (1999). "The Web as a Graph: Measurements, Models, and Methods". Proceedings of the International Conference on Combinatorics and Computing , Tokyo , Japan, July 26 – 28. London: Springer, P. 1 - 17. Available online: http://www.tomkinshome.com/papers/archive/cocoon99.pdf [Accessed on Oct. 2005]

- Najork,M. & Wiener, J.L. (2001). "Breadth – First Crawling Yields High Quality Pages". Proceedings of the 10th World Wide Web Conference (p.114 - 118). Hongkong. May 1 - 5 . New York: ACM. Available online: http://www10.org/cdrom/papers/208/ [Accessed on Oct. 2005]

- Pant, G., Srinivasan, p. Menczer,F. (2004). "Crawling the Web". Web Dynamics. Springer. Availableonline: http://mia.ece.uic.edu/~papers/MediaBot/pdf00001.pdf [Accessed on Oct. 2005]

- Thelwal, Mike (2002). "Methodologies for Crawler Based Web Surveys". Internet Research. 12(2), P. 124 – 138. Available online: www.scms.rgu.ac.uk/staff/fh/CM1008/documents/lecture3.pdf [Accessed on Oct. 2005]

- Yu, Clement & Meng, Weiyi (2004). "Web Search Technology". The Internet Encyclopedia. Hoboken, NJ: Wiley. P 738 – 753.