

## خزنده و ساختواره وب

شعله ارسطوبور<sup>۱</sup>

### چکیده

وب به عنوان بستر فعالیت موتورهای جستجو، ساختاری نموداری دارد. این ساختار حرکت خزنده‌ها در موتورهای جستجو را به روشهایی منطبق بر خود محدود می‌سازد. مقاله حاضر، به بررسی تأثیر ساختار وب بر چگونگی حرکت خزنده‌ها و فعالیت نمایه سازها در موتورهای جستجو می‌پردازد. پس از بحثی مقدماتی در باب نمودارهای جهت دار و فرایند کار خزنده، عمده‌ترین روشهای حرکت خزنده در سطح وب شامل حرکت «عمق - شروع»، «توزیع - شروع» و «بهترین - 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 عمیق - شروع، حرکت توزیع - شروع، حرکت بهترین - شروع، نمایه‌سازی وب

### مقدمه

وب، مجموعه‌ای عظیم از مدارک است که هر یک برای پاسخگویی به یک نیاز بالقوه ایجاد و در بستر اینترنت منتشر گردیده است. حجم عظیم صفحات و اطلاعات موجود در وب و لزوم وجود ابزارهایی جهت سازماندهی دست کم گوشه‌ای از این مجموعه تاکنون به حد کفایت مورد بحث قرار گرفته است و اینکه موتورهای جستجو با

---

۱. دانشجوی کارشناسی ارشد کتابداری و اطلاع‌رسانی دانشگاه فردوسی مشهد

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 شبکه‌ای از اتصالات با یکدیگر مرتبط می‌سازد و از آنجا که ساختار آن هیچ محدودیتی بر چگونگی نشر صفحات، قالب انتشار، و یا چگونگی برقراری اتصالات و تعداد صفحات وب اعمال نمی‌نماید، گسترش چشمگیری یافته است. اما این مسئله هرگز بدین معنا نیست که وب بدون ساختار است. بسیاری از پژوهشها بیانگر وجود روابطی مشخص و منطقی در ساختار وب هستند.

(Barabasi & Albert, ۱۹۹۹; Kleinberg et al., ۱۹۹۹; Albert et al., ۱۹۹۹)

هر یک از این تحقیقات به نوبه خود به نتایج متفاوتی رسیده‌اند. فصل مشترک آنها مسئله ناهماهنگی توزیع صفحات وب است به گونه‌ای که آلبرت و دیگران<sup>۱</sup> (۱۹۹۹) از وب به عنوان یک شبکه جهانی کوچک یاد می‌کنند در حالی که برادر و همکاران<sup>۲</sup> (۲۰۰۰) توزیع وب را پایونی شکل توصیف می‌کنند. بر این اساس، وب از یک هسته که بیشترین اتصالات به آن برقرار شده و دو بال که با توجه به جهت اتصالات، 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 داده‌اند و آن نموداری بودن ساختار وب<sup>۳</sup> است. با توجه به دستاوردهای پژوهشها، ساختار وب را به یقین می‌توان به صورت یک نمودار عظیم ترسیم نمود؛ نموداری که در برگیرنده گره‌ها و اتصالات متعدد است. در این نمودار، صفحات وب همان گره‌ها و اتصالات همان لینکهای برقرار شده میان صفحات مختلف است (Albert et al., ۱۹۹۹; Thelwall, ۲۰۰۲; Evans & Walker, ۲۰۰۴; Yu & Meng, ۲۰۰۴; Coathy, ۲۰۰۴).

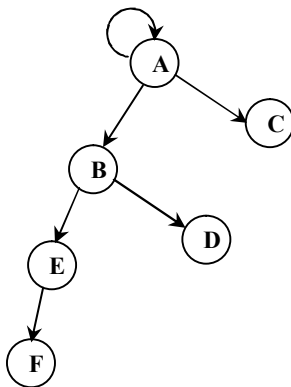
بنابراین وب نه تنها بدون ساختار نیست بلکه می‌توان آن را به صورت یک ساختواره نموداری نیز ترسیم نمود. این ساختواره خواسته یا ناخواسته بر چگونگی گردآوری صفحات و نمایه‌سازی آنها توسط موتورهای جستجو تأثیرگذار بوده و در نهایت نتایج جستجو و میزان ربط آنها را رقم می‌زند.

۱. Albert et al.  
۲. Broder et al.  
۳. Web's Graph Structure

### نمودارهای جهت‌دار<sup>۱</sup>

نمودار در واقع مجموعه‌ای از گره‌ها و خطوط است که آن را به صورت ریاضی  $G(V, E)$  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  $\bar{i}$  نمایش دهیم، برای هر گره که به عنوان گره اصلی در نظر گرفته شود، رابطه  $\bar{i} \in [1, n)$  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 ۱).



تصویر ۱. نمونه‌ای از یک نمودار جهت‌دار

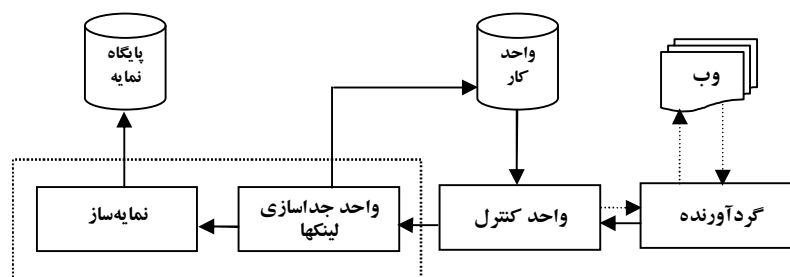
پیشتر بیان شد که هر صفحه از وب دارای تعدادی اتصال است. این اتصالها یا از صفحه هسته<sup>۲</sup> به سایر صفحات و یا از سایر صفحات به صفحه موردنظر برقرار شده است. بنابراین، تمام ویژگیهای پیش گفته، در باب صفحات وب و لینکهای آنها نیز صدق می‌کند. با توجه به جهت برقراری لینکها، می‌توان وب را نیز به صورت یک نمودار جهت‌دار ترسیم نمود.

۱. Directed Graph  
۲. Seed Page

## خزنده

تقریباً تمام خزنده‌ها در موتورهای جستجو دارای چهار بخش گردآورنده<sup>۱</sup>، واحد کنترل<sup>۲</sup>، واحد سازه‌یابی<sup>۳</sup> و واحد کار<sup>۴</sup> هستند. گردآورنده تحت نظارت واحد کنترل، به صفحات هسته (گره‌های مختلف) رفته و مدارک را به واحد جداسازی لینکها می‌فرستد. پس از جداسازی، لینکهای مناسب به واحد کار ارسال شده و در فهرست دستور کار بعدی گردآورنده قرار می‌گیرند. واحد سازه‌یابی در واقع از دو بخش جداسازی لینکها<sup>۵</sup> و نمایه‌سازی تشکیل شده است.

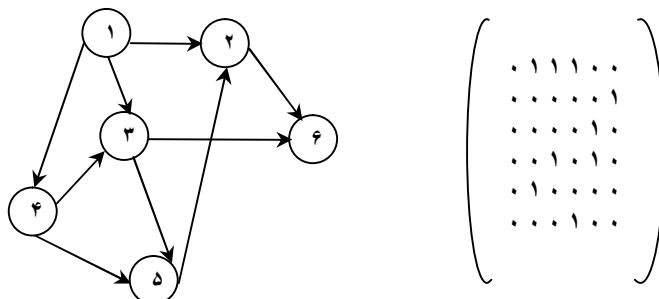
آنچه نهایتاً در پایگاه ذخیره می‌شود در واقع حاصل فرایند نمایه‌سازی است که تحت قالب تعریف شده در الگوریتم موتور جستجو به صورت واژگان و عبارات مختلف در آمده است (Cothey, ۲۰۰۴). تصویر ۲ ساختار یک خزنده را نشان می‌دهد.



تصویر ۲. 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 جبری، حرکت روی نمودارهای جهت‌دار را می‌توان به صورت ماتریسهای حرکتی<sup>۶</sup> ساده کرد. در موتورهای جستجو نیز، با توجه به شباهت ساختاری، استفاده از قواعد ریاضی حاکم بر این ماتریسها، تحلیل و مقایسه را آسان می‌نماید. (تصویر ۳)

۱. Fetcher  
 ۲. Controller  
 ۳. Parsing Unit  
 ۴. Workload Unit  
 ۵. Link Extracting  
 ۶. Traversal Matrix

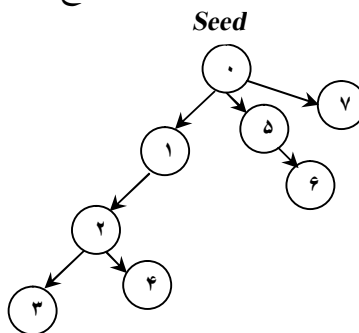


تصویر ۳. ساختواره نموداری وب و ماتریس حرکتی متناظر با جهت‌های مشخص شده

به طور کلی، سه روش برای حرکت خزنده در شبکه لینکهای وب وجود دارد. این سه روش عبارتند از حرکت عمق - شروع<sup>۱</sup>، توزیع - شروع<sup>۲</sup>، و بهترین - شروع<sup>۳</sup>.

### حرکت عمق - 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 - شروع، مانند تصویر ۴ خواهد بود.



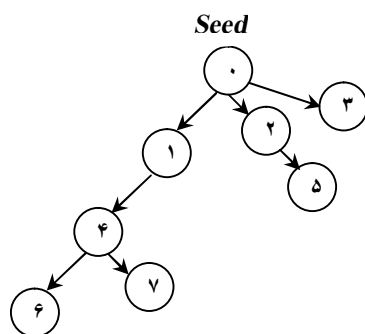
تصویر ۴. صفحه هسته و ترتیب حرکت خزنده با استفاده از الگوریتم عمق - شروع

- ۱. Depth - First
- ۲. Breadth - First
- ۳. Best - First

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 تصویر ۴، چنانچه سطح عمق تعریف شده ۲ باشد، ترتیب حرکت گردآورنده ۱، ۲، ۵، ۶ و ۷ بوده و عملاً صفحه‌های ۳ و ۴ مورد بررسی قرار نخواهند گرفت.

### حرکت توزیع - 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 می‌دهد. به عنوان نمونه، ترتیب حرکت گردآورنده تحت نظارت واحد کنترل و با استفاده از الگوریتم توزیع - شروع در صفحه‌های مختلف مانند تصویر شماره ۵ خواهد بود.



تصویر ۵. صفحه هسته و ترتیب حرکت خزنده با استفاده از الگوریتم توزیع - شروع  
 روش حرکت توزیع - شروع مورد علاقه بسیاری از طراحان برنامه‌های خزنده در موتورهای جستجو است، زیرا طراحی و اجرای آن به صورت رایانه‌ای بسیار ساده‌تر از روش حرکت عمق - 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., ۲۰۰۲).

حال، چنانچه صفحه هسته به یک مطلب خاص پردازد، با توجه به آنکه گردآورنده تمام لینکهای موجود در صفحه و یا صفحات بعد را دنبال نمی کند، حرکت خزنده تأثیر بسیار زیادی بر نمایه سازی و در نهایت بازایی اطلاعات خواهد داشت.

در حرکت عمق - 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 نمایه ساز فرستاده می شوند به احتمال، اغلب حول یک مطلب یا موضوع خواهند بود. در حالی که در حرکت توزیع - شروع گرایش واحد کنترل به حرکت در سطح است و لذا گردآورنده ابتدا به گره های تعیین شده در سطح سرکشی خواهد کرد. در چنین شرایطی صفحاتی برای واحد نمایه ساز 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 می شوند.

بنابراین، با حرکت از سطح به عمق، دید نمایه ساز خواه نا خواه جزء نگر بوده و با حرکت در سطح دید نمایه سازی در حول یک مطلب با گستره ای وسیع تر و جامع تر متمرکز خواهد بود (Herrman, ۲۰۰۳).

با توجه به مطالعات دهه ۱۹۹۰، 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 بازایی بهتر موتورهای جستجوی مختلف در سطح وب نداشت.

تصمیم گیری در باب انتخاب هر یک از این دو روش و تردیدهای موجود، روش سوم را پیش پای طراحان و برنامه نویسان قرار داد و آن، حرکت «بهترین - شروع» بود.

### حرکت بهترین - شروع

بهترین در حوزه حرکت خزنده روی ساختواره وب، در واقع معانی متفاوتی دارد. الگوریتمهای مختلفی برای حرکت بهترین - شروع وجود دارند که بر اساس فرمول محاسبه

بهترین صفحه بعدی، اسامی متفاوت دارند. از این بین می‌توان به خزنده متمرکز<sup>۱</sup>، جستجوی کوسه‌ای<sup>۲</sup>، عنکبوت‌های اطلاعاتی<sup>۳</sup> و... اشاره کرد.

در ساده‌ترین حالت، از سیاست‌های رتبه‌بندی همچون "رتبه‌بندی صفحات"<sup>۴</sup> 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 می‌فرستد.

حرکت بهترین - 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 - 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،  $F_{ii}$  نشانگر صفحاتی که از U به آنها لینک برقرار شده<sup>۵</sup> و  $B_{ii}$  نشانگر صفحاتی باشد که به U لینک برقرار کرده‌اند<sup>۶</sup>

---

۱. Focused Crawler  
۲. Shark Search  
۳. Info Spiders  
۴. Page Rank  
۵. Forward Links  
۶. Back Links

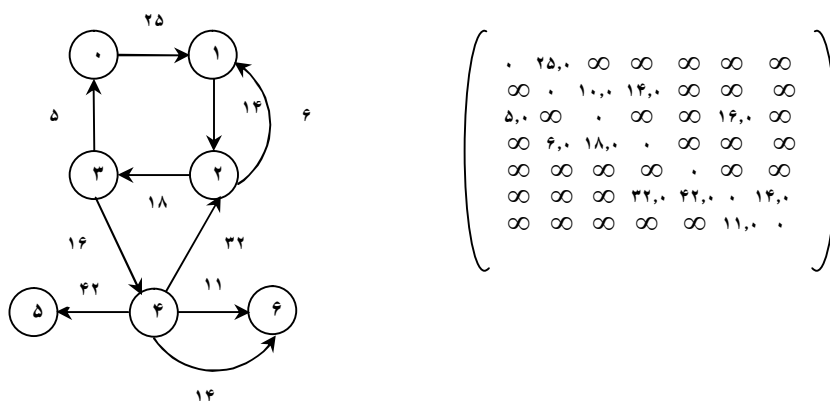


رتبه صفحه از طریق فرمول حاضر و یا فرمولهایی با عناصر اصلی مشابه، قابل سنجش خواهد بود.

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

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<sub>u</sub>| در مخرج، نشانگر وجود رابطه معکوس رتبه U و تعداد لینکهایی است که به صفحات دیگر برقرار کرده است!

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 می‌گردد. تصویر ۶ نشانگر این ترتیب بر اساس رتبه‌های مشخص شده در ماتریس حرکتی خواهد بود.



۱. این مسئله تابع این اصل است که هرچه تعداد صفحات زیرمجموعه صفحه هسته (Child pages) بیشتر باشد، ارزش و رتبه صفحه هسته بین این صفحات تقسیم می‌شود (Yu & Meng, ۲۰۰۴).

### تصویر ۶. لینکها و ارزش عددی هر یک و نمایش آنها در قالب ماتریس حرکتی

این مسئله در نظر کاملاً منطقی است، اما ناجورک و وینر<sup>۱</sup> (۲۰۰۱) 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 بر مبنای 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 - 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<sup>۲</sup> است. پس از آن، بر اساس دستور آن زبان خاص، سازه‌یاب به تعیین ساختار ترکیبی آن رشته می‌پردازد.<sup>۳</sup>

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 می‌دهد. از آنجا که سازه‌یاب بر اساس دستور زبان از قبل

۱. Najork & Wiener

۲. Input Symbols' String

۳. پاکروان، امیرحسین (۱۳۷۶)، فرهنگ کامپیوتر یادواره "انگلیسی-فارسی"، (تهران: یادواره اسدی؛ فرهنگستان یادواره).

تعریف شده به هدف دستیابی به محتوای مشخصی عمل می‌کند، تقسیم‌بندی واژگان استخراج شده و وزن دهی به آنها کار ساده‌ای خواهد بود (Fischer, ۲۰۰۵)

زبان HTML زبان غالب در سطح وب به شمار می‌آید، لذا کلید موتورهای جستجو دارای نرم افزارهای سازه‌یابی سازگار با HTML برای زبانهای مختلف هستند. به واسطه این نرم افزارها، برجسبهای HTML و ارزش آنها به سرعت شناسایی می‌شوند. برای جداسازی و یکسان نمودن آدرسهای موجود در یک صفحه، از سازه‌یابها به منظور شناسایی برجسبهای 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 دستورالعملهایی برای تبدیل حروف بزرگ آدرسها به حروف کوچک، برداشتن انشعابهای اضافی از دنباله آدرس، تصحیح و یا تکمیل برخی از آدرسها و ...، برای سازه‌یابهای مختلف تعریف می‌شود (Pant, Sirinivan & Meczer, ۲۰۰۴).

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 کلید واژگان هم ریشه را در یک محل دارند. در نهایت، وظیفه هر سازه‌یاب ایجاد درخت سازه‌یابی<sup>۱</sup> است. در این مرحله، واحد سازه‌یابی آدرس و یا واژه موجود در صفحه را با کمک محتوا و محل برجسب ارزیابی کرده، درختواره‌ای از ساختار صفحه تشکیل می‌دهد. نمونه‌ای از این درخت و قالب HTML متناظر با آن، در تصویر ۷ نشان داده شده است.

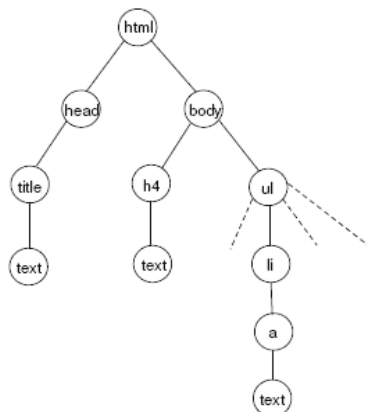
گره اول در این نمودار نشانگر قالب مدرک، گره‌های میانی برجسبهای مختلف مدرک و آخرین گره‌ها نماینده محتوای میان دو برجسب ابتدایی و انتهایی است. تشکیل این درخت، کار وزن‌دهی به هر متن و واژه و عبارت استخراج شده از آن را ساده می‌نماید. رتبه متون در قسمت‌های مختلف صفحه با توجه به الگوریتم رتبه‌بندی

---

۱. Parse Tree

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.

```
<html>
<head>
<title>Projects</title>
</head>
<body>
<h4>Projects</h4>
<ul>
<li> <a href="blink.html">LAMP</a> Linkage analysis with multiple processors.</li>
<li> <a href="nice.html">NICE</a> The network infrastructure for combinatorial exploration.</li>
<li> <a href="amass.html">AMASS</a> A DNA sequence assembly algorithm.</li>
<li> <a href="dali.html">DALI</a> A distributed, adaptive, first-order logic theorem prover.</li>
</ul>
</body>
</html>
```



تصویر ۷. درخت سازه‌یابی و قالب HTML متناظر با آن

### جمع‌بندی

با وجود ماهیت متغیر وب، باز هم ساختار وب بر نحوه سازماندهی آن تأثیرگذار خواهد بود. با توجه به اینکه هم اکنون موتورهای جستجو از مهمترین سازمان‌دهندگان به شمار می‌روند و از طرفی با در نظر گرفتن اینکه ساختار وب روشهای متفاوت و نه متعدد گردآوری اطلاعات را به طراحان نرم افزارهای خزنده دیکته می‌کند، می‌توان بیان داشت که ساختار 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 می‌دهد. تصمیم‌گیری در باب انتخاب شیوه حرکت خزنده - اعم از حرکت به عمق یا حرکت در سطح و یا انتخاب هر صفحه

بسته به کیفیت آن - 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 سازه یاب - چه کوتاه و مختصر (سازه‌یاب‌های ساده) و چه پیچیده (سازه‌یاب‌های سطح بالا) - 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 واسطه شیوه حرکت در سطح وب گردآوری شده اند و نمایه‌سازی روی ساختار ساختمان 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., Jeung, H. & Barabasi, A. (۱۹۹۹). "The Diameter of the World Wide Web". Nature . Vol. ۴۰۱, P. ۱۳۰ Available online: [www.10.org/cdrom/papers/۲۰۸](http://www.10.org/cdrom/papers/۲۰۸) [Accessed on Oct. ۲۰۰۵]

- Barabasi A.L. & Albert, R. (۱۹۹۹). "Emergence of Scaling in random Networks". Science. Vol. ۲۸۶, P ۵۰۹ - ۵۱۲. Available online: <http://www.nd.edu/~networks> [Accessed on Oct. ۲۰۰۵]

- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, s., Stata, R., (۲۰۰۰). "Graph Structure in the Web". Computer Networks. Vol. ۳۳, P. ۳۰۹ - ۳۲۰. Available online: <http://www.9.org/w9cdrom/۱۶۰/۱۶۰.html> [Accessed on Oct. ۲۰۰۵]

- Chakrabarti, s. , Joshi, M.M., Punera, K. & Pennock, D.M. (۲۰۰۲). "The Structure of Broad Topics On the Web". Proceedings of the ۱۱<sup>th</sup> World Wide Web Conference (p.۵۰۸ - ۵۱۰). Honolulu, Hawaii, May ۷- ۱۱ . New York: ACM. Available online: <http://http.cs.berkeley.edu/~soumen/doc/www۲۰۰۲t/p۳۳۸-chakrabarti.pdf> [Accessed on Oct. ۲۰۰۵]

- Cothey, Viv (۲۰۰۴). "Web Crawling reliability". Journal of the American Society for Information Science and Technology. ۵۵(۱۴). P. ۱۲۲۸ - ۱۲۳۸.

- Evans, Michael P. & Walker, Andrew (۲۰۰۴). "Using The Web Graph to Influence Application Behavior". Internet Research. ۱۴(۵). P. ۳۷۲ - ۳۷۸.

- Fischer, Hendrik (۲۰۰۵). Decisions To Go: An Intelligent Mobile Decision Support System[Dissertation]. Georgia: The University of Georgia. Available online: <http://graduate.gradsch.uga.edu/etdarchive/summer۲۰۰۵/fischer-hendrik-۲۰۰۵۰۸-ms.pdf> [Accessed on Oct. ۲۰۰۵]

- Herrmann, Frank (۲۰۰۳). Web search engines. Available online: <http://graduate.gradsch.uga.edu/etdarchive/summer۲۰۰۵/fischer-hendrik-۲۰۰۵۰۸-ms.pdf> [Accessed on Oct. ۲۰۰۵]

- Kleinberg, J., Kumar, R., Raghava, P., Rajagopalan, S., & Tomkins, A. (۱۹۹۹). "The Web as a Graph: Measurements, Models, and Methods". Proceedings of the International Conference on Combinatorics and Computing , Tokyo , Japan, July ۲۶ - ۲۸. London: Springer, P. ۱ - ۱۷. Available online: <http://www.tomkinshome.com/papers/archive/cocoon۹۹.pdf> [Accessed on Oct. ۲۰۰۵]

- Najork, M. & Wiener, J.L. (۲۰۰۱). "Breadth - First Crawling Yields High Quality Pages". Proceedings of the ۱۰<sup>th</sup> World Wide Web Conference (p. ۱۱۴ - ۱۱۸). Hongkong. May ۱ - ۵. New York: ACM. Available online: <http://www.۱۰.org/cdrom/papers/۲۰۰۱/> [Accessed on Oct. ۲۰۰۵]

- Pant, G., Srinivasan, p. Menczer, F. (۲۰۰۴). "Crawling the Web". Web Dynamics. Springer. Available online: <http://mia.ece.uic.edu/~papers/MediaBot/pdf۰۰۰۱.pdf> [Accessed on Oct. ۲۰۰۵]

- Thelwal, Mike (۲۰۰۲). "Methodologies for Crawler Based Web Surveys". Internet Research. ۱۲(۲), P. ۱۲۴ - ۱۳۸. Available online: [www.scms.rgu.ac.uk/staff/fh/CM۱۰۰۸/documents/lecture۳.pdf](http://www.scms.rgu.ac.uk/staff/fh/CM۱۰۰۸/documents/lecture۳.pdf) [Accessed on Oct. ۲۰۰۵]

امکان سنجی ارائه خدمات اشاعه اطلاعات گزینشی ... / ۱۰۹

---

- Yu, Clement & Meng, Weiyi (۲۰۰۴). "Web Search Technology". The Internet Encyclopedia. Hoboken, NJ: Wiley. P ۷۳۸ – ۷۵۳.